Submitted to *Mathematics of Operations Research*
manuscript (Please, provide the mansucript number!)

Optimal Information Blending with Measurements in the L2 Sphere

Boris Defourny
Department of Operations Research and Financial Engineering, Princeton University, Princeton, New Jersey 08544,
defourny@princeton.edu, http://www.princeton.edu/~defourny

Ilya O. Ryzhov
Department of Decision, Operations and Information Technologies, Robert H. Smith School of Business, University of Maryland, College Park, Maryland 20742,
iryzhov@rhsmith.umd.edu, http://www.rhsmith.umd.edu/faculty/iryzhov

Warren B. Powell
Department of Operations Research and Financial Engineering, Princeton University, Princeton, New Jersey 08544,
powell@princeton.edu, http://castlelab.princeton.edu/

We consider a sequential information collection problem where a risk-averse decision-maker updates a Bayesian belief about the unknown objective function of a linear program. The information is collected in the form of a linear combination of the objective coefficients, subject to random noise. We have the ability to choose the weights in the linear combination, creating a new, nonconvex continuous optimization problem, which we refer to as information blending. We develop two optimal blending strategies: an active learning method that maximizes uncertainty reduction, and an economic approach that maximizes an expected improvement criterion. Semidefinite programming relaxations are used to create efficient convex approximations to the nonconvex blending problem.

Key words: Stochastic programming; semidefinite programming; value of information; Markov decision process; risk aversion
MSC2000 subject classification: Primary: 90C22; secondary: 90C40
OR/MS subject classification: Primary: Programming, stochastic
History:

---

1. Introduction. Consider planning problems that can be reformulated as linear programs (LPs) in standard form:

$$\text{maximize } c^\top x \quad \text{subject to } Ax = b,\ x \succeq 0\ . \tag{1}$$

Suppose, however, that the vector of objective coefficients is unknown, and is modeled as a random vector following some multivariate probability distribution. Problems where c is random are well studied in the field of stochastic optimization, covering applications such as production problems with unknown profit margins, or logistics and network problems with uncertain costs.

There are several standard techniques for converting (1) with random c into a well-defined and tractable optimization problem. For instance, the problem $\max \mathbb{E}\{c^\top x\}$ over $x \in \mathcal{X}$, that is,

$$\max_{x\in\mathcal{X}} \bar{c}^\top x\,, \qquad \text{where } \mathcal{X} = \{x \in \mathbb{R}^n : Ax = b, x \succeq 0\}\ , \quad \bar{c} = \mathbb{E}\{c\}\ , \tag{2}$$

optimizes the original objective function $c^\top x$ in expectation. This approach may be reasonable when the decision-maker is risk-neutral with respect to performance. However, in many applications, the decision-maker is likely to be risk-averse, preferring to sacrifice some performance on average while hedging against worst-case scenarios. In such situations, robust optimization [5] offers a way to obtain computationally tractable, conservative decisions. Typically, one would infer a bounded uncertainty set $\mathcal{C}$ with good geometric properties from the distribution of c, and then optimize the worst-case bilinear objective $\max_{x\in\mathcal{X}} \min_{c\in\mathcal{C}} c^\top x$. For instance, suppose that c follows a multivariate normal distribution with covariance matrix Σ, an assumption that will hold throughout the paper:

$$c \sim \mathcal{N}(\bar{c}, \Sigma) \ . \tag{3}$$

As we show in this paper, under some assumptions on the choice of $\mathcal{C}$, we can reformulate the worst-case maximization as the second-order cone program (SOCP) [2]

$$\max_{x\in\mathcal{X}} \bar{c}^\top x - \alpha\sqrt{x^\top \Sigma x}, \tag{4}$$

for some $\alpha > 0$. This problem is polynomially solvable to a fixed precision by interior-point methods. Note that, if $\alpha = 0$, the robust formulation (4) reduces to the risk-neutral formulation (2).

It is possible to reinterpret the distribution in (3) as a Bayesian prior representing the decision-maker's subjective beliefs about c. To emphasize this interpretation, we use the notation c^{true} to represent the unknown vector of objective coefficients, indicating that the decision-maker wishes to estimate some unknown "true" value. The multivariate normal prior $\mathcal{N}(\bar{c}, \Sigma)$ is a convenient way to incorporate correlations in the decision-maker's beliefs about the unknown c^{true}. Correlations reflect a belief about the similarity of the unknown coefficients. For example, the unknown profit margins for two similar products can reasonably be assumed to be correlated.

With this interpretation, we consider situations where the decision-maker has the ability to collect additional information about c^{true} before implementing a solution $x \in \mathcal{X}$ in production. A single piece of information about c^{true} will change the parameters of the belief distribution (3), thus changing the optimal solution of (4). Simply put, the uncertainty set from which we draw the worst-case scenario is now itself subject to change. Moreover, if we have multiple opportunities to collect information, we face a new problem of optimal multi-stage information collection. In this problem, the goal is to guide the evolution of the uncertainty set in a way that improves the performance of the robust solution to

$$v_\alpha(\bar{c}, \Sigma) = \max_{x:\ Ax=b,\ x\geq 0} \bar{c}^\top x - \alpha\sqrt{x^\top \Sigma x}. \tag{5}$$

We seek to develop sequential, adaptive information collection policies with the ability to learn from the outcomes of previous observations.

The work by [51] studies the information collection problem in the context of the risk-neutral decision made in (2), under the assumption that the decision-maker collects information in the form of scalar noisy measurements of individual objective coefficients c_j^{true}. The present study adds the dimension of risk-averse decision-making, and makes the following major generalization: we study systems where information on the full vector $c^{\text{true}} \in \mathbb{R}^n$ can be acquired by observing

$$y = u^\top c^{\text{true}} + w, \tag{6}$$

where $u \in \mathbb{R}^n$ is a *measurement vector* chosen in the ball $\mathbb{B} = \{u \in \mathbb{R}^n : ||u||_2 \leq 1\}$, $w \sim \mathcal{N}(0, \sigma_w^2)$ is an independent Gaussian noise of variance $\sigma_w^2 > 0$, and y is the noisy observation that depends on

the measurement vector. Given the observation y, by Bayesian updating, c follows the posterior distribution $\mathcal{N}(\bar{c}', \Sigma')$ given by

$$\bar{c}' = \bar{c} + \frac{\Sigma u}{u^\top \Sigma u + \sigma_w^2}(y - \bar{c}^\top u), \tag{7}$$

$$\Sigma' = \Sigma - \frac{\Sigma u u^\top \Sigma}{u^\top \Sigma u + \sigma_w^2} \ , \tag{8}$$

and the optimal value $v_\alpha(\bar{c}, \Sigma)$ in (5) is updated to $v_\alpha(\bar{c}', \Sigma')$. Instead of requiring measurements of individual objective coefficients, which can be done by restricting $u = e_j$ where e_j is the jth unit vector in $\mathbb{R}^n$, we allow a "blended" observation that provides information about multiple unknown values simultaneously. See Section 2 below for a discussion of applications that motivate this approach to learning.

With these elements in place, the objective of the problem can be formally stated as follows. Let $\{u_k : k \geq 1\}$ be a sequence of measurement vectors, and let $\{\bar{c}_k\}$, $\{\Sigma_k\}$ be the belief parameters obtained by recursively applying (7)-(8) to obtain $\bar{c}_{k+1}, \Sigma_{k+1}$ from $\bar{c}_k$, Σ_k, u_k, and $y_{k+1} = u_k^\top c^{\text{true}} + w_{k+1}$, where the noise terms $\{w_k\}$ are i.i.d. A policy $\pi : \mathbb{R}^n \times \mathbb{R}^{n \times n} \mapsto \mathbb{R}^n$ determines the kth measurement vector $u_k = \pi(\bar{c}_k, \Sigma_k)$ dynamically based on the most current information. The theoretical optimal policy π^* maximizes

$$\sup_\pi \mathbb{E}^\pi v_\alpha(\bar{c}_K, \Sigma_K), \tag{9}$$

where K is a fixed information budget. The objective in (9) is an example of *offline learning*, where a finite period of exploration is followed by a single *implementation decision* at time K. It is important to note that, in (9), the decision-maker is risk-averse with respect to the implementation decision (represented by v_α), but *risk-neutral* with respect to the outcomes of the measurements (represented by $\mathbb{E}^\pi$). This reflects the fact that, in many applications of offline learning, the cost of a poor measurement is much less than the cost of a poor implementation. For example, if simulation is used to collect information, clearly the cost of a poor or uninformative simulation is much less than the cost of a poor implementation in the field. The work by [49] provides additional theoretical justification for this model; essentially, risk-aversion with respect to measurements leads to overly conservative policies that do not learn enough about the problem.

Equation (9) describes a multi-stage stochastic dynamic programming problem with a multi-dimensional and continuous state space [22]. This problem is intractable, motivating the development of other policies that are suboptimal for (9), but may be optimal with respect to other relevant and more tractable criteria. In this paper, we study two such policies. First, we analytically derive a policy that achieves the optimal rate of uncertainty reduction in our beliefs about c^{true}. We show that this policy chooses u to be a dominant eigenvector of the posterior covariance matrix of c at each time step. Second, we develop a different policy that trades uncertainty reduction against the performance of the robust solution in (5) using the expected improvement criterion:

$$\mathbb{K}_\alpha(u, \bar{c}, \Sigma) = \mathbb{E}_y\{v_\alpha(\bar{c}', \Sigma') \mid u, \bar{c}, \Sigma\} - v_\alpha(\bar{c}, \Sigma), \tag{10}$$

$$u^* \in \arg\max_{u \in \mathbb{B}} \mathbb{K}_\alpha(u, \bar{c}, \Sigma). \tag{11}$$

The problem (11) takes the nonlinear update equations (7,8) into account. Inside the expectation, there is a change of optimal x for each outcome y, so as to obtain $v_\alpha(\bar{c}', \Sigma')$. It is easy to see that such a policy is optimal for (9) if $K = 1$, and we also demonstrate that the same policy is asymptotically optimal as $K \to \infty$, suggesting that it may also work well for finite time horizons.

Although (11) defines a nonconvex optimization problem, we develop computationally tractable convex relaxations that reformulate (11) as a semidefinite program. We then show numerically that

the SDP relaxation enables us to find directions u that achieve higher expected improvement than the unit-vector policy of [51]. The key theoretical insight of these results is that the information content of a scalar observation can be significantly improved by optimally blending information, instead of observing individual problem parameters.

The paper is organized as follows. Section 2 discusses related work. Section 3 derives the robust objective (5) from the definition of uncertainty sets for c. Section 4 applies the framework to Markov decision processes. Section 5 establishes properties of optimal solutions for the measurement selection problem (11). Section 6 studies the measurement policy that maximizes the rate of uncertainty reduction. Section 7 presents the main results of the paper on the optimization of (11). Section 8 discusses the asymptotic convergence of the expected improvement policy. Section 9 presents numerical work, and Section 10 concludes.

2. Context and related work.

This section provides additional context for our study. First, in Section 2.1, we discuss applications that motivate our key modeling choices. Second, in Section 2.2, we discuss related theoretical and methodological work from optimal learning.

2.1. Motivating applications.

Our model has three important distinguishing features: 1) an offline objective, where the information collection process is separated from the final implementation decision at time K; 2) a risk-averse implementation decision, expressed as the solution to a second-order cone program; 3) blended observations, where information takes the form of a linear combination of the unknown parameters, rather than a noisy observation of an individual parameters. We now discuss two classes of applications where these features may be needed to address key problem characteristics.

First, the model of a linear program with random coefficients can be applied to characterize the optimal policy solving a finite-state Markov decision process (MDP) [43], where randomness in c corresponds to the situation of a one-period reward function that is not perfectly known. Recent work that has specifically considered MDPs with known transition probabilities, but unknown reward functions, includes [34, 57, 44]. Specific applications include problems in artificial intelligence where the state is the position of a robotic agent, and the reward is based on the agent's environment (e.g. searching for an object or detecting a source of radiation). The transition probabilities are deterministic since the agent moves in accordance with user commands, but the reward function is unknown due to the agent's uncertainty about the environment. This example is discussed by [34]. Other examples of uncertainty in the reward function may include cases where the reward function comes from human judgment (e.g. utility elicitation) or is estimated empirically (as in reinforcement learning).

Second, our model is applicable to problems involving *learning in linear regression*. The updating equations (7)-(8) are identical to those used in Bayesian linear regression [35], in which the unknown values of a large set of alternatives are related to a small set of parameters via a linear regression model. For example, [36] considers the following problem in drug discovery. A base molecule is modified by adding atoms to various sites on the molecule. The number of possible configurations of atoms grows combinatorially in the number of sites; however, a standard approach is to model the value of a configuration as a linear combination of the values of individual sites. We can then perform a laboratory experiment with a chosen configuration. The outcome will be a scalar quantity, used to update our beliefs about the regression features.

However, the set of alternatives available for implementation may not be finite. For a recent application where this might not be the case, see the recent work by [7]. This work empirically estimates a linear regression model that predicts the effectiveness of a cancer treatment (in terms of patient survival metrics) using features such as the instantaneous dose and average dose of various component drugs. The treatment is then optimized by choosing the doses that maximize

survival subject to linear constraints on the dosage levels. Robust optimization is suggested as a way of hedging against worst-case treatment outcomes. The regression model is estimated using the results of previous clinical trials, available in the medical literature.

Such applications exhibit a feedback loop between statistics and optimization. First, we learn the coefficients of a linear regression model. Then, we solve a continuous robust optimization problem to determine the best *regression features* (dosages) to use in practice, based on coefficients estimated from clinical trials. Our model, studied in this paper, provides a way to integrate these two stages: in practice, data from previous clinical trials can be used to guide the design of new trials, and the data from these new trials will in turn change our beliefs about the regression coefficients. We can now use (7)-(8) to update our beliefs after every new observation, and we can also use (10)-(11) to guide the design of the next clinical trial.

Another recent application of learning in linear regression is given in [29]. This work estimates a linear regression model that predicts the success rate of a non-profit fundraising campaign as a linear function of various design choices, such as the format or the subject matter of a letter mailed to the donors. Fundraisers are conducted regularly (e.g. monthly), and the outcome of a campaign can be used to improve the design of the next campaign. Although the number of configurations for a campaign is finite, it is combinatorially large, leading to an application of the SDP relaxation techniques originally developed in this paper.

Thus, our work has the potential to contribute to applications of analytics, where incoming data are used to guide high-impact decisions with significant penalties for worst-case outcomes. We design policies that recommend information to collect based on the potential of this information to improve the quality of the implementation decision.

2.2. Literature review. The present paper brings together two different streams of literature. The first is the work on robust optimization [5, 8]. Robust solutions to linear programs with uncertainty have been extensively studied [6], and the theory of robust optimization has also been developed for Markov decision processes [38, 30, 45]. See also [48] for recent work connecting the robust solution and the uncertainty set to a risk measure chosen by the decision-maker. Particularly relevant to the present paper is [16], which derived an expression of the form (5) applied specifically to MDPs. However, the notion that sequential information collection may change the uncertainty set over time, thus also changing the robust solution, has received much less attention. To give an example, equation (10) for measurement selection in robust MDPs was previously stated in [15] for $u \in \{e_1, \ldots, e_n\}$; however, the computational approach in this study was based on an approximation that did not take into account the change of the optimal solution from $\arg\max_x v_\alpha(\bar{c}, \Sigma)$ to $\arg\max_x v_\alpha(\bar{c}', \Sigma')$. In Section 4, we discuss this approach in more detail, in the perspective of contrasting it with our new results, which use SDP relaxations to approximate (10) more closely, while also allowing information blending.

The second major stream is the literature on statistical learning and sequential information collection, usually known in different communities by the name of a particular problem. Examples include ranking and selection in simulation [32], multi-armed bandits in applied probability [24] and computer science [3], and global optimization [31]. This paper is closest to the simulation perspective, in which the information collection process ("ranking") is usually separated from the final implementation decision ("selection"). Such models typically consider applications where information is obtained from simulation models, or from limited field experiments, for a finite length of time. After this stage is complete, a single alternative is selected and implemented in the field. In this literature, the implementation decision usually consists of choosing the largest value in a finite set; by contrast, our model is closer to [50, 51], where the ranking and selection framework is generalized to allow implementation decisions that optimize a mathematical program with unknown parameters. We also adopt the Bayesian framework for information collection; see [11] or [42] for a survey of Bayesian learning methods in simulation optimization.

The multi-armed bandit literature has also considered similar problems from the point of view of online learning, in which the measurement and implementation decisions are the same. That is, an experiment is made by implementing an alternative, and the resulting observation also serves as an economic reward. The objective is then to maximize the cumulative reward across all experiments, rather than the value of a final implementation. Recent work in this area has considered problem variants that allow information blending [47, 14], as well as risk-averse performance measures [9]. There are two primary differences between this work and the setting in our paper. First, we consider an offline problem in which measurement is separated from implementation. The offline setting has substantial structural differences from the bandit setting: for example, an index policy is optimal for an online problem [24], but not an offline problem. Second, although we model information as a linear function of u, the problem that we are learning about is no longer linear, but rather is a SOCP obtained by transforming the robust optimization problem. In the language of this community, our problem can be described as "offline SOCP with linear feedback."

We study two policies for choosing the information blend u. Our first policy maximizes the rate of uncertainty reduction achieved by each measurement. This approach is along the lines of active learning in statistics [13], where the objective is to minimize uncertainty (i.e. improve the accuracy of a statistical model), with no regard for the economic value of a set of estimates. Conversely, our second policy is based on the expected improvement criterion, previously developed by [31] for global optimization and [28] for ranking and selection. This approach, also known by the names "value of information" [11] or "knowledge gradient" [22], provides an economic valuation of information in terms of the average improvement contributed by a single measurement to the optimal value of (2) or (4). This computation balances the expected value of the current solution to (2) or (4) against the decision-maker's uncertainty about that solution (and therefore the potential to improve it).

In the simulation literature, the decision-maker is almost always assumed to be risk-neutral [12], and the expected improvement criterion is defined in terms of the risk-neutral problem (2). Recently, however, there has been some interest in integrating concepts of risk-aversion and robust optimization into simulation optimization [56, 17]. To our knowledge, the work by [49] is the first to formally link ranking and selection with robust optimization, using a model that is risk-neutral with respect to information, but risk-averse with respect to implementation. The present paper also adopts this approach, and the formulation in (5) covers both risk-neutral ($\alpha = 0$) and risk-averse ($\alpha > 0$) implementation decisions.

To summarize, our work contributes to the literature on sequential learning as well as robust optimization. We show how two types of optimal information blends can be computed via an SDP reformulation, which also enables us to improve on a heuristic previously developed for robust Markov decision processes.

3. Robust optimization criterion.

In statistics, confidence intervals can describe uncertain scalar parameters. The intervals are often mean-centered, although nonsymmetric choices are possible. The width of the interval is chosen to achieve a given confidence level $1-\epsilon$. For $c \sim \mathcal{N}(\bar{c}, \Sigma)$ with Σ positive definite ($\Sigma \succ 0$), we consider for some $\alpha > 0$ the confidence ellipsoid

$$\mathcal{C} = \{c \in \mathbb{R}^n : (c-\bar{c})^\top \Sigma^{-1}(c-\bar{c}) \leq \alpha^2\}. \tag{12}$$

Choosing $\alpha^2 = F^{-1}_{\chi^2_n}(1-\epsilon)$, where $F^{-1}_{\chi^2_n}(\cdot)$ is the inverse cumulative distribution function (cdf) of the chi-square distribution with n degrees of freedom, ensures that $c \in \mathcal{C}$ with probability $1-\epsilon$.

By selecting $\mathcal{C}$ as the uncertainty set for c, tractable robust optimization programs can be obtained.

Lemma 1. *With* $\mathcal{X} = \{x \in \mathbb{R}^n : Ax = b, x \succeq 0\}$ *and* $\mathcal{C}$ *given by* (12), *the problem* $\max_{x\in\mathcal{X}} \min_{\tilde{c}\in\mathcal{C}} \tilde{c}^\top x$ *is equivalent to* $\max_{x\in\mathcal{X}} \bar{c}^\top x - \alpha\sqrt{x^\top \Sigma x}$.

Proof. If $\alpha = 0$, $\mathcal{C} = \{\bar{c}\}$ and the result is trivially verified. If $\alpha > 0$, for any fixed x, the inner minimum $\min_{\tilde{c}\in\mathcal{C}} \tilde{c}^\top x$ is computed by applying the change of variable $z = \Sigma^{-1/2}(\tilde{c} - \bar{c})$, which yields $\min_{z:z^\top z \le \alpha^2} (\Sigma^{1/2} z + \bar{c})^\top x$ where $\bar{c}^\top x$ is fixed. The minimum is attained at $z^* = -\beta\Sigma^{1/2}x$ for β such that $||z^*||_2^2 = \alpha^2$, that is, $\beta = \alpha/\sqrt{x^\top \Sigma x}$. In terms of $\tilde{c}$ the optimal solution is $\tilde{c} = \bar{c} - \alpha\Sigma x/\sqrt{x^\top\Sigma x}$, hence the value for the inner minimum, $\bar{c}^\top x - \alpha\sqrt{x^\top\Sigma x}$. □

If Σ is only positive semidefinite ($\Sigma \succeq 0$ but $\Sigma \not\succ 0$), we consider the confidence ellipsoid

$$\begin{aligned} \widetilde{\mathcal{C}} &= \{c = Q_0 Q_0^\top \bar{c} + Q_+ c_+ \in \mathbb{R}^n : c_+ \in \mathcal{C}_+\} \\ \mathcal{C}_+ &= \{c_+ \in \mathbb{R}^p : (c_+ - Q_+^\top \bar{c})^\top \Sigma_+^{-1} (c_+ - Q_+^\top \bar{c}) \le \alpha^2\}, \end{aligned} \tag{13}$$

where $Q_+ \in \mathbb{R}^{n\times p}$ and $Q_0 \in \mathbb{R}^{n\times(n-p)}$ come from the singular value decomposition (svd)

$$\Sigma = QSQ^\top = [Q_+\ Q_0] \begin{bmatrix} \Sigma_+ & 0 \\ 0 & 0 \end{bmatrix} [Q_+\ Q_0]^\top, \tag{14}$$

Σ_+ being the diagonal matrix containing the p positive singular values of Σ.

Lemma 2. *With* $\mathcal{X} = \{x \in \mathbb{R}^n : Ax = b, x \succeq 0\}$ *and* $\widetilde{\mathcal{C}}$ *given by* (13), *the problem* $\max_{x\in\mathcal{X}} \min_{c\in\widetilde{\mathcal{C}}} c^\top x$ *is equivalent to* $\max_{x\in\mathcal{X}} \bar{c}^\top x - \alpha\sqrt{x^\top\Sigma x}$.

Proof. Using (14), c can be reexpressed as

$$c = Q_0 Q_0^\top \bar{c} + Q_+ c_+, \qquad\qquad c_+ \sim \mathcal{N}(Q_+^\top \bar{c}, \Sigma_+).$$

Then,

$$\begin{aligned} \max_{x\in\mathcal{X}} \min_{c\in\mathcal{C}} c^\top x &= \max_{x\in\mathcal{X}} \min_{c_+\in\mathcal{C}_+} (Q_0 Q_0^\top \bar{c} + Q_+ c_+)^\top x \\ &= \max_{x\in\mathcal{X}} \{\bar{c}^\top Q_0 Q_0^\top x + \min_{c_+\in\mathcal{C}_+} c_+^\top (Q_+^\top x)\} \\ &= \max_{x\in\mathcal{X}} \bar{c}^\top Q_0 Q_0^\top x + (Q_+^\top \bar{c})^\top (Q_+^\top x) - \alpha\sqrt{(Q_+^\top x)^\top \Sigma_+ Q_+^\top x} \\ &= \max_{x\in\mathcal{X}} \bar{c}^\top (Q_0 Q_0^\top + Q_+ Q_+^\top) x - \alpha\sqrt{x^\top Q_+ \Sigma_+ Q_+^\top x}, \end{aligned}$$

which reduces to $\max_{x\in\mathcal{X}} \bar{c}^\top x - \alpha\sqrt{x^\top\Sigma x}$ using (14) and $Q_0 Q_0^\top + Q_+ Q_+^\top = [Q_0\ Q_+][Q_0\ Q_+]^\top = Q^\top Q = I$. □

Recall that we have adopted a Bayesian approach to the estimation of c^{true}. We assume that the belief about c^{true}, expressed by $c^{\text{true}} \sim \mathcal{N}(\bar{x}, \Sigma)$, is correct. It follows that for any $x \in \mathcal{X}$, the belief on the quantity $x^\top c^{\text{true}}$ is expressed by $x^\top c^{\text{true}} \sim \mathcal{N}(x^\top \bar{c}, x^\top \Sigma x)$. In particular, for a solution $\bar{x}$ to $\max_{x\in\mathcal{X}} \bar{c}^\top x - \alpha\sqrt{x^\top\Sigma x}$, the belief on the quantity $\bar{x}^\top c^{\text{true}}$ is expressed by $\bar{x}^\top c^{\text{true}} \sim \mathcal{N}(\bar{x}^\top \bar{c}, \bar{x}^\top \Sigma \bar{x})$. Now, by definition of $v_\alpha(\bar{c}, \Sigma)$, we have $v_\alpha(\bar{c}, \Sigma) = \bar{x}^\top \bar{c} - \alpha\sqrt{\bar{x}^\top\Sigma\bar{x}}$, whence

$$\begin{aligned} \mathbb{P}\{\bar{x}^\top c^{\text{true}} \ge v_\alpha(\bar{c}, \Sigma)\} &= \mathbb{P}\{\bar{x}^\top c^{\text{true}} \ge \bar{x}^\top \bar{c} - \alpha\sqrt{\bar{x}^\top\Sigma\bar{x}}\} \\ &= \mathbb{P}\left\{\frac{\bar{x}^\top c^{\text{true}} - \bar{x}^\top \bar{c}}{\sqrt{\bar{x}^\top\Sigma\bar{x}}} \ge -\alpha\right\} = \Phi(\alpha), \end{aligned}$$

where Φ is the cumulative distribution function (cdf) of $\mathcal{N}(0,1)$. Thus if we want to ensure with confidence $1-\epsilon$ that $\bar{x}^\top c^{\text{true}} \ge v_\alpha(\bar{c}, \Sigma)$, we can choose $\alpha = \Phi^{-1}(1-\epsilon)$, which is less conservative than the choice $\alpha^2 = F^{-1}_{\chi^2_n}(1-\epsilon)$.

Finally, let us mention that (5) can be solved as a quadratic program with quadratic constraints (QCQP):

LEMMA 3. *If $\Sigma \succ 0$, a dual formulation to (5) is*

$$v_\alpha(\bar{c},\Sigma) = \min_{c,z} b^\top z \;\text{ subject to } c \in \mathcal{C},\; A^\top z \succeq c \;\;,$$

using $\mathcal{C}$ given by (12). Otherwise, using $\widetilde{\mathcal{C}}$ given by (13),

$$v_\alpha(\bar{c},\Sigma) = \min_{c_+,z} b^\top z \;\text{ subject to } c_+ \in \mathcal{C}_+,\;\; A^\top z \succeq Q_0 Q_0^\top \bar{c} + Q_+ c_+ \;\;.$$

Proof. A dual problem to $\max_{x\in\mathcal{X}} \bar{c}^\top x - \alpha\sqrt{x^\top \Sigma x}$ or equivalently $\max_{x\in\mathcal{X}} \min_{c\in\mathcal{C}} c^\top x$ is $\min_{c\in\mathcal{C}} \max_{x\in\mathcal{X}} c^\top x$, relying on the fact that $\mathcal{X}$ and $\mathcal{C}$ are nonempty compact convex sets. The dual to $\max_{x\in\mathcal{X}} c^\top x$ is $\min_{z\in\mathcal{Z}} b^\top z$ for $\mathcal{Z} = \{\mathbb{R}^m : A^\top z \succeq c\}$, hence the overall problem. The version with $\widetilde{\mathcal{C}}$ can be established similarly. □

4. Application to Markov decision processes. Let the tuple (S, A, P, R) define a Markov decision process [43] where S is a finite state space with $|S|$ states, A is a finite action space with $|A|$ actions, $P : S \times A \times S \mapsto [0,1]$ with values $p(s'|s,a)$ is a transition probability function, and $R : S \times A \mapsto \mathbb{R}$ is a reward function with bounded values $r(s,a)$. Let $0 < \gamma < 1$ be a discount factor, and let $b(j) = \mathbb{P}\{s_0 = j\}$ specify an initial state distribution, states being labeled from 1 to $|S|$. The maximization of the expected discounted cumulated reward

$$v_\gamma^\pi = \mathbb{E}^\pi\{\textstyle\sum_{t=0}^\infty \gamma^t r(s_t,a_t)\} \tag{15}$$

by the choice of a stochastic policy $\pi : S \times A \mapsto [0,1]$ with values $\pi(s,a) = \mathbb{P}\{a_t = a | s_t = s\}$ admits a dual linear programming formulation [18]

$$\begin{aligned} \text{maximize} \quad & \sum_{s\in S}\sum_{a\in A} r(s,a)x(s,a) \\ \text{subject to} \quad & \sum_{a\in A} x(j,a) - \sum_{s\in S}\sum_{a\in A} \gamma\, p(j|s,a)x(s,a) = b(j) \qquad \text{for } j \in S, \\ & x(s,a) \geq 0 \qquad \text{for } s \in S, a \in A, \end{aligned} \tag{16}$$

which is of the form (1). Given an optimal $x^* \in \mathbb{R}^{|S|\times|A|}$,

$$\pi^*(s,a) = x^*(s,a) / \textstyle\sum_{a'\in A} x^*(s,a') \tag{17}$$

is an optimal stochastic policy. The optimization variables $x(s,a)$ (occupation measures) represent the total discounted probability of being in state s and choosing action a, when the system starts from state j with probability $b(j)$. The optimal policy (17) will be independent of the initial distribution.

4.1. MDP with Bayesian prior. In our framework, we assume that the rewards $r(s,a)$ are unknown but endowed with a prior $\mathcal{N}(\bar{r},\Sigma)$, where $\bar{r}$ collects the means $\bar{r}(s,a)$ and Σ is the covariance matrix collecting elements $\Sigma(s,a;s',a')$. The framework is less general than (Bayesian, model-based) reinforcement learning (RL), where transition probabilities would also be endowed with a prior. Nonetheless, the framework is already a valuable step for studying model ambiguity in Markov decision processes from a Bayesian standpoint.

Under the risk-neutral approach ($\alpha = 0$), the rewards $r(s,a)$ in (16) are set to their Bayesian mean $\bar{r}(s,a)$. The optimization problem has still the structure of a MDP, implying the existence of an optimal deterministic policy. To see that from (16), note that the simplex algorithm returns a vertex solution x^* defined by $|S| \cdot |A|$ linear equations, $|S|$ coming from the equality constraints

and $|S| \cdot |A| - |S|$ coming from active inequalities $x(s,a) = 0$. Hence x^* has at most $|S|$ nonzero coordinates. The definition of a proper policy requires one nonzero coordinate being assigned to each state, implying that the policy (17) is in fact deterministic.

When the robust optimization approach is used ($\alpha > 0$), the program for finding an optimal policy becomes

$$\begin{aligned} \text{maximize} \quad & \sum_{s\in S}\sum_{a\in A} \bar{r}(s,a)x(s,a) - \alpha\sqrt{\sum_{s\in S}\sum_{a\in A}\sum_{s'\in S}\sum_{a'\in A} x(s,a)\Sigma(s,a;s',a')x(s',a')} \qquad (18)\\ \text{subject to} \quad & \sum_{a\in A} x(j,a) - \sum_{s\in S}\sum_{a\in A} \gamma\, p(j|s,a)x(s,a) = b(j) \qquad \text{for } j \in S,\\ & x(s,a) \geq 0 \qquad \text{for } s \in S,\ a \in A. \end{aligned}$$

Generically, optimal solutions to SOCPs are not vertex solutions. Thus more elements $x^*(s,a)$ will be nonzero, and the resulting stochastic policy (17) does not necessarily degenerate into a deterministic one.

The program (18) is a tractable robust MDP obtained by applying generic robust linear programming techniques. The covariance matrix $\Sigma(s,a;s'a')$ allows one to model worst-case reward dependencies among state-action pairs.

4.2. Optimal measurements with fixed decisions. Consider now the measurement selection problem based on the maximization over u of $\mathbb{K}_\alpha(u,\bar{c},\Sigma)$ as defined by (10). An approximation proposed in [15] for robust MDPs with the measurement u valued in $\{e_1,\dots,e_n\}$ assumes that, inside the expectation in (10), for each outcome y, the optimal solution x' attaining $v_\alpha(\bar{c}',\Sigma')$ is replaced by the solution $\bar{x}$ attaining $v_\alpha(\bar{c},\Sigma)$. By doing so, $\mathbb{K}_\alpha(u,\bar{c},\Sigma)$ is approximated by

$$\widetilde{\mathbb{K}}_\alpha(u,\bar{c},\Sigma) = \mathbb{E}_y\{[\bar{c}'^\top\bar{x} - \alpha\sqrt{\bar{x}^\top\Sigma'\bar{x}}] - [\bar{c}^\top\bar{x} - \alpha\sqrt{\bar{x}^\top\Sigma\bar{x}}] \mid u,\bar{c},\Sigma\} = \alpha(\sqrt{\bar{x}^\top\Sigma\bar{x}} - \sqrt{\bar{x}^\top\Sigma'\bar{x}}) \ , \qquad (19)$$

where $\mathbb{E}_y\{\bar{c}'\} = \bar{c}$ has been used.

Note that $\widetilde{\mathbb{K}}_\alpha(u,\bar{c},\Sigma) = 0$ for all u if $\alpha = 0$, suggesting that this approximation is uninformative in the risk-neutral case. Despite this undesirable behavior, we can still investigate the problem of maximizing $\widetilde{\mathbb{K}}_\alpha(u,\bar{c},\Sigma)$.

PROPOSITION 1. *Let $\bar{x} \in \arg\max_{x\in\mathcal{X}} \bar{c}^\top x - \alpha\sqrt{x^\top\Sigma x}$, and let $\widetilde{\mathbb{K}}_\alpha(\cdot,\bar{c},\Sigma)$ be the approximation relative to $\bar{x}$. Then either $\Sigma\bar{x} = 0$ and any $u \in \mathbb{B}$ is optimal for $\max_{u\in\mathbb{B}} \widetilde{\mathbb{K}}_\alpha(\cdot,\bar{c},\Sigma)$, or $\Sigma x \neq 0$ and the maximum of $\widetilde{\mathbb{K}}_\alpha(\cdot,\bar{c},\Sigma)$ over $\mathbb{B}$ is attained by selecting*

$$\bar{u} \in \left\{ \pm\frac{(\Sigma + \mathrm{I}_n\sigma_w^2)^{-1}\Sigma\bar{x}}{||(\Sigma + \mathrm{I}_n\sigma_w^2)^{-1}\Sigma\bar{x}||} \right\} \ , \qquad (20)$$

where I_n is the identity matrix in $\mathbb{R}^{n\times n}$.

Proof. The proof relies on techniques used in Section 5. It can be found in the appendix. □

From (20), we can better understand the effect of the fixed-decision approximation. If we assume momentarily that σ_w^2 is small with respect to the eigenvalues of Σ, then $(\Sigma + \mathrm{I}_n\sigma_w^2)^{-1}\Sigma$ is close to I_n, so $\bar{u}$ is close to $\bar{x}/||\bar{x}||$. Therefore, $\bar{u}$ tends to measure the coordinates of c^{true} according to the magnitude of their believed contribution to the objective value given the current optimal solution $\bar{x}$. For any value of σ_w^2, if Σ is diagonal, the coordinates c_j for $j \in \{i : \bar{x}_i = 0\}$ are not measured.

This analysis suggests that using the approximation (19) would lead to a measurement policy that is not asymptotically consistent, in the sense that wrong beliefs would not necessarily be corrected by an infinite sequence of measurements.

5. Structural properties for optimal measurements. Convex functions have their supremum on the boundary of their effective domain [46]. A similar result holds for the nonconvex function $\mathbb{K}_\alpha(\cdot, \bar{c}, \Sigma)$.

THEOREM 1. *Let $\mathcal{U}$ be an arbitrary nonempty closed convex bounded set. Let $\partial\mathcal{U}$ denote the boundary of $\mathcal{U}$. We have*

$$\max_{u\in\mathcal{U}} \mathbb{K}_\alpha(u, \bar{c}, \Sigma) = \max_{u\in\partial\mathcal{U}} \mathbb{K}_\alpha(u, \bar{c}, \Sigma).$$

Proof. Fix u in the interior of $\mathcal{U}$. Define u_+ by extending u to $\partial\mathcal{U}$ as follows: define $t^* = \max\{t \geq 0 : tu/||u|| \in \mathcal{U}\}$, $\tau = t^*/||u||$, $u_+ = \tau u \in \partial\mathcal{U}$. Necessarily, $\tau \geq 1$. Essentially, we show that the measurement based on u_+ dominates the measurement based on u, so that optimal measurements are on $\partial\mathcal{U}$.

Define

$$\beta = \frac{u^\top \Sigma u + \sigma_w^2}{u^\top \Sigma u + (\sigma_w/\tau)^2}, \quad \Lambda = \frac{\Sigma u u^\top \Sigma}{u^\top \Sigma u + \sigma_w^2}.$$

Note that $1 \leq \beta \leq \tau^2$. From the update of Σ after measurements $y_u = c^\top u + w$ or $y_{u_+} = c^\top u_+ + w$, we deduce the ordering of the two updated covariance matrices in the cone of the positive semidefinite matrices:

$$\Sigma'_{u^+} = \Sigma - \frac{\Sigma u_+ u_+^\top \Sigma}{u_+^\top \Sigma u_+ + \sigma_w^2} = \Sigma - \frac{\tau^2 \Sigma u u^\top \Sigma}{\tau^2[u^\top \Sigma u + (\sigma_w/\tau)^2]} = \Sigma - \beta\Lambda \preceq \Sigma - \Lambda = \Sigma'_u,$$

meaning (informally) that the residual uncertainty is "smaller" with u_+. From the update of $\bar{c}$ after the observations y_u or y_{u_+},

$$\bar{c}'_u = \bar{c} + \frac{\Sigma u}{u^\top \Sigma u + \sigma_w^2}(y_u - \bar{c}^\top u), \qquad \bar{c}'_{u_+} = \bar{c} + \frac{\Sigma u_+}{u_+^\top \Sigma u_+ + \sigma_w^2}(y_{u_+} - \bar{c}^\top u),$$

and from the distribution of the observations,

$$y_u \sim \mathcal{N}(u^\top \bar{c}, u^\top \Sigma u + \sigma_w^2), \qquad y_{u_+} \sim \mathcal{N}(\tau u^\top \bar{c}, \tau^2 u^\top \Sigma u + \sigma_w^2),$$

we deduce the distribution of the updated means,

$$\bar{c}'_u \sim \mathcal{N}(\bar{c}, \Lambda), \qquad \bar{c}'_{u_+} \sim \mathcal{N}(\bar{c}, \beta\Lambda).$$

Using the zero-mean random vector $z \sim \mathcal{N}(0, \Lambda)$, we have

$$\mathbb{E}\{v_\alpha(\bar{c}'_{u_+}, \Sigma'_{u_+})\} = \mathbb{E}\{v_\alpha(\bar{c} + \sqrt{\beta}z, \Sigma'_{u_+})\} \geq \mathbb{E}\{v_\alpha(\bar{c} + z, \Sigma'_{u_+})\} = \mathbb{E}\{v_\alpha(\bar{c}'_u, \Sigma'_{u_+})\},$$

where the inequality is justified by an extension of Jensen's inequality, that states that a function $g(t) = \mathbb{E}\{f(x_0 + tz)\}$ defined for $t \geq 0$ is monotone increasing if f is convex and $\mathbb{E}\{z\} = 0$. Since $\Sigma'_{u_+} \preceq \Sigma'_u$, we have

$$\bar{c}'^\top_u x - \alpha\sqrt{x^\top \Sigma'_{u_+} x} \geq \bar{c}'^\top_u x - \alpha\sqrt{x^\top \Sigma'_u x},$$

implying $v_\alpha(\bar{c}'_u, \Sigma'_{u_+}) \geq v_\alpha(\bar{c}'_u, \Sigma'_u)$ and thus

$$\mathbb{E}\{v_\alpha(\bar{c}'_u, \Sigma'_{u_+})\} \geq \mathbb{E}\{v_\alpha(c'_u, \Sigma'_u)\}.$$

Therefore, $\mathbb{K}(u^+, \bar{c}, \Sigma) \geq \mathbb{K}(u, \bar{c}, \Sigma)$. Since u was arbitrary, the result follows. □

If we now restrict ourselves to the case where $\mathcal{U}$ is the L2-ball $\mathbb{B}$, Theorem 1 indicates that we should seek solutions u on the L2-sphere $\partial\mathbb{B} = \{u \in \mathbb{R}^n : ||u|| = 1\}$.

It will be convenient to rewrite the objective (10) as

$$\mathbb{K}_\alpha(u, \bar{c}, \Sigma) = \mathbb{E}_t\{v_\alpha(\bar{c} + t\,\Sigma d_u, \Sigma') \;\mid\; u, \bar{c}, \Sigma\} - v_\alpha(\bar{c}, \Sigma) \;\;, \tag{21}$$

where $t \sim \mathcal{N}(0,1)$ and where we have introduced the vector

$$d_u = \frac{u}{\sqrt{u^\top \Sigma u + \sigma_w^2}} \;\;. \tag{22}$$

In the special case $||u|| = 1$, we have $u^\top \Sigma u + \sigma_w^2 = u^\top(\Sigma + \sigma_w^2 \mathrm{I}_n)u$. This leads us to define

$$P = \Sigma + \sigma_w^2 \mathrm{I}_n \;\;. \tag{23}$$

The matrix P is positive definite and thus invertible.

In the risk-neutral case ($\alpha = 0$), we can go further in the characterization of optimal solutions.

THEOREM 2. *Assume the risk-neutral case ($\alpha = 0$). Then either any $u \in \mathbb{B}$ is optimal for $\max_{u\in\mathbb{B}} \mathbb{K}_0(u, \bar{c}, \Sigma)$, or the solutions u^* optimal for $\max_{u\in\mathbb{B}} \mathbb{K}_0(u, \bar{c}, \Sigma)$ satisfy*

$$u^* \in \left\{ \pm \frac{P^{-1}\Sigma\,\mathbb{E}\{t\bar{x}(t)\}}{||P^{-1}\Sigma\,\mathbb{E}\{t\bar{x}(t)\}||} \right\} \;\;, \qquad \bar{x}(t) \in \arg\max_{x\in\mathcal{X}} \left(\bar{c} + \frac{t\,\Sigma u^*}{||P^{1/2}u^*||} \right)^\top x \;\;,$$

where the expectation is taken over $t \sim \mathcal{N}(0,1)$, and where without loss of generality the vector-valued function $\bar{x}(\cdot)$ is piecewise-constant on $\mathbb{R}$ with a finite number of pieces.

Proof. Let Ξ denote the space of all measurable vector-valued functions $x(\cdot) : \mathbb{R} \mapsto \mathbb{R}^n$ with values $x(t) \in \mathcal{X}$, defined for all $t \in \mathbb{R}$. Note first that for any $u \in \mathbb{B}$, there exists for each t a selection $x(t)$ [19] of the optimal solution set $X(t) = \arg\max_{x\in\mathcal{X}} (\bar{c} + t\Sigma d_u)^\top x$ such that $x(\cdot) \in \Xi$ is a piecewise-constant, vector-valued function with a finite number of pieces [23, 51]. Thus we can actually restrict Ξ to that space of functions. Consider

$$\begin{aligned}
\max_{u\in\mathbb{B}} \mathbb{K}_0(u, \bar{c}, \Sigma) &= \max_{u\in\mathbb{B}} \mathbb{E}_t \left\{ \max_{x(\cdot)\in\Xi} \left(\bar{c} + t\frac{\Sigma u}{\sqrt{u^\top \Sigma u + \sigma_w^2}} \right)^\top x(t) \right\} - v_0(\bar{c}, \Sigma) \\
&= \max_{x(\cdot)\in\Xi} \max_{u\in\mathbb{B}} \mathbb{E}_t \left\{ \left(\bar{c} + t\frac{\Sigma u}{\sqrt{u^\top \Sigma u + \sigma_w^2}} \right)^\top x(t) \right\} - v_0(\bar{c}, \Sigma) \;\;,
\end{aligned}$$

where the interchange between $\mathbb{E}_t$ and $\max_{x(\cdot)\in\Xi}$ is possible because the optimization problem is written in terms of a function $x(\cdot)$ that does not explicitly depend on u.

One can check that $\max_{u\in\mathbb{B}} \mathbb{K}_0(u, \bar{c}, \Sigma) \geq 0$ by plugging in the constant-valued function $x(\cdot) = \bar{x}_0$, where $\bar{x}_0 \in \mathcal{X}$ attains $v_0(\bar{c}, \Sigma)$: for any u, one obtains $\mathbb{E}_t\{(\bar{c} + t\Sigma d_u)^\top \bar{x}_0\} = \bar{c}^\top \bar{x}_0 + \mathbb{E}\{t\} d_u^\top \Sigma \bar{x}_0 = v_0(\bar{c}, \Sigma)$.

Assume that we are given an optimal function $\bar{x}(\cdot) \in \Xi$ for the problem. The set of the vectors $u \in \mathbb{B}$ that attain $\max_{u\in\mathbb{B}} \mathbb{K}_0(u, \bar{c}, \Sigma)$ along with $\bar{x}(\cdot)$ can be expressed by

$$\arg\max_{u\in\mathbb{B}} \mathbb{E} \left\{ \left(\bar{c} + t\,\frac{\Sigma u}{\sqrt{u^\top \Sigma u + \sigma_w^2}} \right)^\top \bar{x}(t) \right\} = \arg\max_{u\in\mathbb{B}} \frac{u^\top}{\sqrt{u^\top \Sigma u + \sigma_w^2}} \Sigma \mathbb{E}\left\{ t\bar{x}(t) \right\} \;\;,$$

dropping the constant term $\mathbb{E}\{\bar{c}^\top \bar{x}(t)\}$ on the right-hand side.

If $\Sigma\mathbb{E}\{t\bar{x}(t)\}=0$, then any $u\in\mathbb{B}$ is optimal. Otherwise, $\Sigma\mathbb{E}\{t\bar{x}(t)\}\neq 0$, and by theorem 1,

$$\arg\max_{u\in\mathbb{B}}\mathbb{K}_0(u,\bar{c},\Sigma)=\arg\max_{u:\,||u||=1}\frac{u^\top\Sigma\mathbb{E}\{t\bar{x}(t)\}}{\sqrt{u^\top Pu}}.$$

Moreover, using $v=P^{1/2}u$, we have

$$\max_{u:\,||u||=1}\frac{u^\top\Sigma\mathbb{E}\{t\bar{x}(t)\}}{\sqrt{u^\top Pu}}=\max_{v:\,||P^{-1/2}v||=1}\frac{v^\top P^{-1/2}\Sigma\mathbb{E}\{t\bar{x}(t)\}}{||v||}\ .$$

Recall that for any z, here taken to be $z=P^{-1/2}\Sigma\mathbb{E}\{t\bar{x}(t)\}$,

$$||z||=\max_{y\in\mathbb{B}}y^\top z=\max_{y\neq 0}y^\top z/||y||\ .$$

Therefore, an optimal v is given by $v^*=\beta P^{-1/2}\Sigma\mathbb{E}\{t\bar{x}(t)\}$ with β such that $||P^{-1/2}v^*||=1$. Then it follows that $u^*=P^{-1/2}v=P^{-1}\Sigma\mathbb{E}\{t\bar{x}(t)\}/||P^{-1}\Sigma\mathbb{E}\{t\bar{x}(t)\}||$ is optimal. Moreover, if u^* is optimal, then $-u^*$ is optimal, by the symmetry of the Gaussian distribution and the expression of d_{u^*}. □

Corollary 1 **(Norm-maximization reformulation)**. *In the risk-neutral case ($\alpha=0$), we have*

$$\max_{u\in\mathbb{B}}\mathbb{K}_0(u,\bar{c},\Sigma)=\max_{x(\cdot):\,x(t)\in\mathcal{X}}\left\{\mathbb{E}_t\{\bar{c}^\top x(t)\}+||P^{-1/2}\Sigma\mathbb{E}_t\{tx(t)\}||\right\}-v_0(\bar{c},\Sigma)\ , \tag{24}$$

where u is recovered from an optimal $x^*(\cdot)$ by $u^*=P^{-1}\Sigma\mathbb{E}_t\{tx^*(t)\}/||P^{-1}\Sigma\mathbb{E}_t\{tx^*(t)\}||$.

Proof. Let $f(x(\cdot))=\mathbb{E}_t\{\bar{c}^\top x(t)\}+||P^{-1/2}\Sigma\mathbb{E}_t\{tx(t)\}||\}$. Since f is convex, optimal solutions are attained on the extreme points of the feasibility set. Thus without loss of generality we can assume that $x(t)$ is a vertex of $\mathcal{X}$ for each t. Let $\bar{x}(\cdot)\in\Xi$ be an optimal solution with Ξ defined as in Theorem 2.

First, consider the degenerate case where $\Sigma\mathbb{E}_t\{t\bar{x}(t)\}=0$. Then, $f(\bar{x}(\cdot))=\mathbb{E}_t\{\bar{c}^\top\bar{x}(t)\}$. Since $\bar{x}(t)$ is optimal by assumption, and since any solution $\bar{x}_0$ that attains $v_0(\bar{c},\Sigma)$ is in $\arg\max_{x\in\mathcal{X}}\bar{c}^\top x$, we can assume without loss of generality that $\bar{x}(t)=\bar{x}_0$ almost surely, so that $\mathbb{E}_t\{\bar{c}^\top\bar{x}(t)\}=\bar{c}^\top\bar{x}_0=v_0(\bar{c},\Sigma)$. Hence in that case any measurement is optimal (in fact no new measurement is needed).

Next, consider the nondegenerate case where $\Sigma\mathbb{E}_t\{t\bar{x}(t)\}\neq 0$. The relation $\max_{u\in\mathbb{B}}\mathbb{K}_0(\bar{c},\Sigma)=\max_{x(\cdot)\in\Xi:x(t)\in\mathcal{X}}f(x(\cdot))-v_0(\bar{c},\Sigma)$ can be checked by comparing the two objectives with u set to $P^{-1}\Sigma\mathbb{E}\{tx(t)\}/||P^{-1}\Sigma\mathbb{E}\{tx(t)\}||$: one gets

$$\mathbb{E}\left\{\left(\bar{c}+t\frac{\Sigma\bar{u}}{\sqrt{\bar{u}^\top\Sigma\bar{u}+\sigma_w^2}}\right)^\top\bar{x}(t)\right\}-v_0(\bar{c},\Sigma)=\bar{c}^\top\mathbb{E}\{\bar{x}(t)\}+||P^{-1/2}\Sigma\mathbb{E}\{t\bar{x}(t)\}||-v_0(\bar{c},\Sigma)\ .$$

At the same time, with $\Sigma\mathbb{E}_t\{t\bar{x}(t)\neq 0$, the subdifferential of $f(x(\cdot))$ at $\bar{x}(\cdot)$ is a singleton corresponding to the gradient of $f(x(\cdot))$ at $\bar{x}(\cdot)$. The gradient of $f(x(\cdot))$ with respect to $x(t')$ for some fixed t' is given by

$$\nabla_{x(t')}f(x(\cdot))=\phi(t')\bar{c}+\phi(t')(t'P^{-1/2}\Sigma)^\top\frac{P^{-1/2}\Sigma\mathbb{E}\{tx(t)\}}{||P^{-1/2}\Sigma\mathbb{E}\{tx(t)\}||}=\phi(t')\left[\bar{c}+t'\frac{\Sigma P^{-1}\Sigma\mathbb{E}\{tx(t)\}}{||P^{-1/2}\Sigma\mathbb{E}\{tx(t)\}||}\right].$$

At $\bar{x}(\cdot)$, we have the implicit definition $\bar{u}=P^{-1}\Sigma\mathbb{E}\{t\bar{x}(t)\}/||P^{-1}\Sigma\mathbb{E}\{t\bar{x}(t)\}||$, so we have

$$\frac{\Sigma\bar{u}}{||P^{1/2}\bar{u}||}=\frac{\Sigma P^{-1}\Sigma\mathbb{E}\{t\bar{x}(t)\}}{||P^{-1/2}\Sigma\mathbb{E}\{t\bar{x}(t)\}||}\ .$$

Therefore, the gradient with respect to $x(t')$ at $\bar{x}(\cdot)$ can be written as

$$\nabla_{x(t')} f(x(\cdot))|_{\bar{x}} = \phi(t') \left[\bar{c} + t' \frac{\Sigma P^{-1} \Sigma \mathbb{E}\{t\bar{x}(t)\}}{||P^{-1/2} \Sigma \mathbb{E}\{t\bar{x}(t)\}||} \right] = \phi(t') \left[\bar{c} + t' \frac{\Sigma \bar{u}}{||P^{1/2}\bar{u}||} \right] \quad .$$

From the basic variational inequality for minimization [19, Thm. 2A.6], a necessary condition for attaining a maximum is $\nabla_{x(t')} f(\bar{x}(\cdot))_{\bar{x}} \in N_{\mathcal{X}}(\bar{x}(t'))$ for almost every t', where $N_{\mathcal{X}}(\bar{x}(t'))$ is the normal cone to $\mathcal{X}$ at $\bar{x}(t')$. Since $\phi(t') > 0$, we can invoke the property that $x \in K$ iff $ax \in K$ for a cone K and some positive scalar a, and deduce that $\bar{x}(\cdot)$ must satisfy

$$\bar{c} + \frac{t\Sigma\bar{u}}{||P^{1/2}\bar{u}||} \in N_{\mathcal{X}}(\bar{x}(t)) \qquad \text{for almost every } t \ .$$

Now, note that these conditions are necessary and sufficient for ensuring that

$$\bar{x}(t) \in \arg\max_{x \in \mathcal{X}} \left(\bar{c} + \frac{t\Sigma\bar{u}}{||P^{1/2}\bar{u}||} \right)^{\top} x \qquad \text{for almost every } t \ ,$$

since the latter problem is convex. We have thus verified that (24) fulfills at optimality the necessary conditions of Theorem 2. □

Theorem 2 and its corollary concern the case $\alpha = 0$ only. They will not be used in the rest of the paper. However, the structure of the problem (24) makes it easier to establish a complexity result:

PROPOSITION 2 **(NP-completeness).** *The decision problem associated with (10) with a discretized expectation is NP-complete.*

Proof. For establishing a complexity result, without loss of generality we can set $\alpha = 0$, $\bar{c} = 0$, $\Sigma = \mathrm{I}_n$, and consider $\max_{u \in \mathbb{B}} \mathbb{K}_0(u, 0, \mathrm{I}_n)$. From (24) we obtain $\max_{x(\cdot):\, x(t) \in \mathcal{X}} (1+\sigma_w^2)^{-1/2} ||\mathbb{E}\{tx(t)\}||$, which is equivalent to $\max_{z \in \mathcal{Z}} ||z||$ with $\mathcal{Z} = \{z \in \mathbb{R}^n : z = \mathbb{E}\{tx(t)\}, x(t) \in \mathcal{X}\}$. By discretizing the random variable t into N samples t_i, we obtain a set $\mathcal{Z}^N$ in $\mathbb{R}^n$ which is the projection of a polyhedral set in $\mathbb{R}^{n(N+1)}$ where each $x(t_i)$ can be assumed to be a vertex of $\mathcal{X}$. In that case, $\mathcal{Z}^N$ is polyhedral. The decision problem associated to the maximization of the L2-norm of a vector over a polyhedral set is known to be NP-complete [33]. □

Proposition 2 indicates that we should not expect to develop exact solution algorithms for our problem. Rather it emphasizes the need for good approximations.

6. Optimal uncertainty reduction. Consider the sequential measurement setting, where measurements are taken iteratively. For a given sequence $\{u_k : k \geq 1\}$ of measurements, let $\Sigma_1 = \Sigma \in \{S \in \mathbb{R}^{n \times n} : S = S^{\top}, S \succeq 0\}$ be the initial covariance matrix, and consider the matrix sequence $\{\Sigma_k : k \geq 1\}$ defined from (8) by

$$\Sigma_{k+1} = \Sigma_k - \Sigma_k u_k u_k^{\top} \Sigma_k / (u_k^{\top} \Sigma_k u_k + \sigma_w^2).$$

Independently of the objective (11) based on the expected value of information from the next measurement, a direct approach for reducing the uncertainty is to acquire information on c^{true} by making measurements u_k such that Σ_k provably tends to the zero matrix. By the degeneracy of the posterior distribution of $c_k \sim \mathcal{N}(\bar{c}_k, \Sigma_k)$, Doob's consistency theorem [20] implies that the sequence of updated means $\bar{c}_k$ tends to c^{true}.

This section studies such a method, and shows that it achieves a rate of convergence which is optimal in a certain sense. Namely, we consider u_k taken as a dominant eigenvector of Σ_k:

$$u_k \in E_{\max}(\Sigma_k), \tag{25}$$

using the following notations defined for any symmetric matrix $S \in \mathbb{R}^{n \times n}$:

- $\lambda_{\max}(S) = \max\{\lambda \in \mathbb{R} : Su = \lambda u,\ u^\top u = 1\}$: largest eigenvalue of S;
- $E_{\max}(S) = \{u \in \mathbb{R}^n : Su = \lambda_{\max}(S)u,\ u^\top u = 1\}$: the set of normalized eigenvectors in the eigenspace associated to $\lambda_{\max}(S)$, excluding the zero vector.

For any $\epsilon > 0$, we can ensure that $\operatorname{trace}\Sigma_k < \epsilon$ after a certain number of measurements, as made precise by the following lemma.

LEMMA 4. *Let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of Σ_1, with repetition according to eigenvalue multiplicity. Fix $\epsilon > 0$. Then the matrix sequence $\{\Sigma_k : k \geq 1\}$ associated with u_k given by (25) satisfies $\operatorname{trace}\Sigma_k < \epsilon$ for any $k > k_0 = \sum_{i=1}^n \log(n/\epsilon)/\log(1/s_i)$, where $s_i = [1 - \lambda_i/(\lambda_i + \sigma_w^2)]$ for $i = 1, \ldots, n$.*

Proof. By the eigenvalue decomposition of $\Sigma_k \succeq 0$, we have $\Sigma_k = \sum_{i=1}^n \lambda_{ik} u_{ik} u_{ik}^\top$, where $\lambda_{1k} \geq \lambda_{2k} \geq \cdots \geq \lambda_{nk} \geq 0$, and where $u_{ik}^\top u_{jk} = 1$ if $i = j$, $u_{ik}^\top u_{jk} = 0$ if $i \neq j$. Taking $u_k = u_{1k}$ in the update equation gives $\Sigma_{k+1} = \Sigma_k - (\lambda_{1k}^2 u_{1k} u_{1k}^\top)/(\lambda_{1k} + \sigma_w^2) = \lambda_{1k}(1 - \lambda_{1k}/(\lambda_{1k} + \sigma_w^2)) u_{1k} u_{1k}^\top + \sum_{i=2}^n \lambda_{ik} u_{ik} u_{ik}^\top$. Therefore, iterations leave the original eigenvectors unchanged.

If the noise variance $\sigma_w^2 = 0$, the covariance would become the zero matrix after at most n iterations (exactly n iterations if the matrix is full rank). With $\sigma_w^2 > 0$, we evaluate the number of iterations needed to have $\operatorname{trace}(\Sigma_k) < \epsilon$ as follows. For each i, let $s_i = 1 - \lambda_{i1}/(\lambda_{i1} + \sigma_w^2)$. Define $k_i = \inf\{k \in \mathbb{N} : s_i^k < \epsilon/n\}$, that is, $k_i = \lceil \log(\epsilon/n)/\log(s_i) \rceil$. Since each iteration shrinks the current largest eigenvalue, we are guaranteed to have $\lambda_{ik} < \epsilon/n$ for each i after $k_0 = \sum_{i=1}^n k_i$ iterations. This implies $\operatorname{trace}\Sigma_k = \sum_{i=1}^n \lambda_{ik} < \epsilon$. □

COROLLARY 2. *The matrix sequence $\{\Sigma_k : k \geq 1\}$ associated to $u_k \in E_{max}(\Sigma_k)$ converges to the zero matrix (in the metric space of the Frobenius norm).*

Proof. $||\Sigma_k||_F = (\sum_{i=1}^n \sum_{j=1}^n \Sigma_{k,ij}^2)^{1/2} = (\sum_{i=1}^n \lambda_{ik}^2)^{1/2} \leq \sum_{i=1}^n |\lambda_{ik}| = \sum_{i=1}^n \lambda_{ik} = \operatorname{trace}(\Sigma_k)$, so $\operatorname{trace}(\Sigma_k) < \epsilon$ implies $||\Sigma_k||_F < \epsilon$. □

Taking new measurements at iterations $k+1$, $k+2$, …, can never increase the uncertainty, in the sense that $\operatorname{trace}(\Sigma_\ell) \leq \operatorname{trace}(\Sigma_k)$ for $\ell > k$. This will hold true for any measurement policy, say π, that maps an information state $(\bar{c}_k, \Sigma_k)$ to some measurement u_k. Now suppose π has some interesting properties, but is not asymptotically consistent. By alternating measurements selected by π and measurements selected by the trace-minimization policy, one can synthetize a new policy which is asymptotically consistent. This observation is summarized in the following corollary.

COROLLARY 3. *Let $\pi : \mathbb{R}^n \times \mathbb{R}^{n \times n} \mapsto \mathbb{R}^n$ denote a measurement policy with values $u_k = \pi(\bar{c}_k, \Sigma_k)$, where k is the iteration counter. Let κ be an integer greater or equal to 2. Let $\pi^\kappa : \mathbb{R}^n \times \mathbb{R}^{n \times n} \times \mathbb{N} \mapsto \mathbb{R}^n$ be a new measurement policy defined by $\pi^\kappa(\bar{c}_k, \Sigma_k, k) = \pi(\bar{c}_k, \Sigma_k)$ if $\mod(k, \kappa) \neq \kappa - 1$, $\pi^\kappa(\bar{c}_k, \Sigma_k, k) \in E_{max}(\Sigma_k)$ if $\mod(k, \kappa) = \kappa - 1$. Then the policy π^κ is asymptotically consistent in the sense that $\operatorname{trace}\Sigma_k < \epsilon$ for any $k > \kappa k_0$, where k_0 is given by Lemma 4.*

Proof. The result follows from the definition of π^κ. □

The following result shows that the rate of convergence cannot be improved.

THEOREM 3. *All the measurement sequences defined by $u_k \in E_{max}(\Sigma_k)$ achieve the optimal rate of convergence of $\{\operatorname{trace}(\Sigma_k) : k \geq 1\}$ to 0, among the sequences such that $||u_k|| \leq 1$.*

Proof. The rate of convergence is maximized if we minimize the trace of Σ_{k+1} given Σ_k. To see why this is true, consider a sequence of M measurements $u_k, \ldots, u_{k+M-1}$. Observe that Σ_{k+M} given Σ_k is invariant under permutations of the measurements. This can be seen from the M rank-one updates of the precision matrix: $[\Sigma_{k+M}]^{-1} = [\Sigma_k]^{-1} + \sum_{\ell=0}^{M-1} u_\ell u_\ell^\top / \sigma_w^2$. Therefore, we don't have to consider postponing a measurement that brings the largest trace reduction, when we jointly optimize over the sequence of measurements.

Writing Σ' for Σ_{k+1} and Σ for Σ_k, we consider

$$\min_{u:\, ||u||=1} \operatorname{trace}\left(\Sigma - \frac{\Sigma u u^\top \Sigma}{u^\top \Sigma u + \sigma_w^2}\right) = \operatorname{trace}(\Sigma) - \max_{u:\, ||u||=1} \frac{u^\top \Sigma \Sigma u}{u^\top P u}.$$

The solution to the maximization problem in the second term is obtained by considering the generalized eigenvalue problem $\Sigma^2 u = \lambda P u$ and taking the vector u associated to the dominant generalized eigenvalue λ. Since P is nonsingular, the generalized eigenvalue problem is equivalent to the standard eigenvalue problem $P^{-1}\Sigma^2 u = \lambda u$. Therefore, the sequence defined by $u_k \in E_{\max}((\Sigma_k + \mathrm{I}_n\sigma_w^2)^{-1}\Sigma_k^2)$ maximizes the rate of convergence of $\mathrm{trace}(\Sigma_k)$ to 0.

We will now prove that $E_{\max}(P^{-1}\Sigma^2) = E_{\max}(\Sigma)$, allowing us to conclude that $u_k \in E_{\max}(\Sigma_k)$ is also optimal. To do that, we use the eigenvalue decomposition $\Sigma = QDQ^\top$, where D is diagonal with elements $D_{ii} = \lambda_i$ such that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$, and $Q = [q_1 \dots q_n]$ is the matrix of eigenvectors such that $Q^\top Q = \mathrm{I}_n = QQ^\top$. By the Rayleigh quotient representation, $u_k \in E_{\max}(P^{-1}\Sigma^2)$ iff $u_k \in \arg\max_{u:\,||u||=1} u^\top P^{-1}\Sigma^2 u$. Now, we have

$$\begin{aligned}
&\arg\max_{u:\,||u||=1} u^\top(\Sigma + \mathrm{I}_n\sigma_w^2)^{-1}\Sigma^2 u \\
=&\arg\max_{u:\,||u||=1} u^\top(Q(D+\mathrm{I}_n\sigma_w^2)Q^\top)^{-1}QD^2Q^\top u = \arg\max_{u:\,||u||=1} u^\top Q(D+\mathrm{I}_n\sigma_w^2)^{-1}D^2Q^\top u \\
=&\arg\max_{\theta:\,||\theta||=1} \theta^\top(D+\mathrm{I}_n\sigma_w^2)^{-1}D^2\theta = \arg\max_{\theta:\,||\theta||=1} \sum_{i=1}^n \frac{\lambda_i^2\theta_i^2}{\lambda_i+\sigma_w^2} = \arg\max_{\theta:\,||\theta||=1} \sum_{i=1}^n \nu_i\theta_i^2,
\end{aligned}$$

where we have used the change of variable $\theta = Q^\top u$ and defined $\nu_i = \lambda_i^2/(\lambda_i + \sigma_w^2)$. We have $\nu_i = \nu_j$ iff $\lambda_i = \lambda_j$. The ordering of the λ_i's implies $\nu_1 \geq \nu_2 \geq \cdots \geq \nu_n \geq 0$. If $\nu_1 > \nu_2$, the optimal solution θ^* is the unit vector e_1, so $u^* = Q\theta^* = Qe_1 = q_1$. If $\nu_1 = \cdots = \nu_k > \nu_{k+1}$, we have $\theta^* \in \{\sum_{i=1}^k w_k e_k : \sum_{i=1}^k w_i = 1,\ w_i \geq 0\}$ and thus $u^* \in \{\sum_{i=1}^k w_k q_k : \sum_{i=1}^k w_i = 1,\ w_i \geq 0\}$, showing that the principal eigenspaces of Σ and $P^{-1}\Sigma^2$ coincide. □

Note that the condition $\Sigma_k \to 0$ is sufficient but not necessary for the convergence of x_k to a maximizer of the true problem (1). To see that, imagine that some coefficient c_j plays no role in the optimization problem, because of a constraint $x_j = 0$. Say that c_j is statistically independent of the other coefficients, and has a prior with an arbitrarily large variance. A sequential measurement algorithm defined by (25) will dedicate many measurements to the reduction of uncertainty on c_j. However, with $\alpha = 0$ we should never measure c_j, since updates of $\bar{c}_j$ never improve the objective.

7. Optimal expected improvement. We now come back to the problem of solving (11) as a stochastic program. A prerequisite is the construction of a finite approximation to the expectation in (10). To do that, consider

- $\phi(t) = (2\pi)^{-1/2}\exp\{-t^2/2\}$: pdf of $\mathcal{N}(0,1)$
- $\Phi(t) = \int_{-\infty}^t \phi(t')dt'$: cdf of $\mathcal{N}(0,1)$
- $\{t_i\}_{1\leq i\leq N}$: sequence defined by $t_0 = -\infty$, $t_{N+1} = +\infty$,

$$\int_{(t_{i-1}+t_i)/2}^{(t_i+t_{i+1})/2} (t - t_i)\phi(t)\mathrm{d}t = 0, \quad 1 \leq i \leq N. \tag{26}$$

The relation (26) expresses a stationary property satisfied by the optimal solution to the quantization problem [25]

$$D_N = \inf_{q\in\mathcal{Q}_N} \mathbb{E}\{||t - q(t)||^2\}, \quad t \sim \mathcal{N}(0,1),$$

where $\mathcal{Q}_N$ denotes the class of measurable functions $q : \mathbb{R} \mapsto \mathbb{R}$ with at most N values $t_1, \dots, t_N$. Because $\mathcal{N}(0,1)$ is one-dimensional and strongly unimodal, the points t_i are uniquely determined by (26) [25, Thm I.5.1]. The points can be computed by methods described in [39].

• $\{p_i\}_{1\le i\le N}$ with $p_i = \Phi\left(\frac{t_i+t_{i+1}}{2}\right) - \Phi\left(\frac{t_{i-1}+t_i}{2}\right)$. For a function f that is Lipschitz continuous modulus L,

$$\left|\mathbb{E}\{f(t)\} - \sum_{i=1}^N p_i f(t_i)\right| \le L\mathbb{E}\{||t-q(t)||\}.$$

For a convex function f, we have [39]

$$\sum_{i=1}^N p_i f(t_i) \le \mathbb{E}\{f(t)\}. \tag{27}$$

Using the optimal N-quantization of $\mathcal{N}(0,1)$, we then define

$$\widehat{\mathbb{K}}_\alpha^N(u,\bar{c},\Sigma) = \sum_{i=1}^N p_i v_\alpha(\bar{c}+t_i\Sigma d_u,\Sigma') - v_\alpha(\bar{c},\Sigma). \tag{28}$$

LEMMA 5. *For all N, $\widehat{\mathbb{K}}_\alpha^N(u,\bar{c},\Sigma) \le \mathbb{K}_\alpha(u,\bar{c},\Sigma)$.*

Proof. For each fixed (x,Σ), the function $\bar{c}^\top x - \alpha\sqrt{x^\top\Sigma x}$ is linear in $\bar{c}$ and thus convex in $\bar{c}$. The maximum over an infinite family of convex functions indexed by x is convex, thus $v_\alpha(\bar{c},\Sigma)$ is convex in $\bar{c}$. Since composition with linear functions preserves convexity, $v_\alpha(\bar{c}+t\Sigma d_u,\Sigma')$ is convex in t. The inequality of the lemma follows from (27). □

Finally, noting that to each $v_\alpha(\bar{c}+t_i\Sigma d_u,\Sigma')$, $i=1,\dots,N$, is associated a program with decision vector $x_i\in\mathbb{R}^n$, and using the update formula for the inverse covariance matrix $[\Sigma']^{-1} = \Sigma^{-1} + uu^\top/\sigma_w^2$, we expand (28) as

$$\widehat{\mathbb{K}}_\alpha^N(u,\bar{c},\Sigma) = \max_{x_1\in\mathcal{X},\dots,x_N\in\mathcal{X}} \sum_{i=1}^N p_i\left[(\bar{c}+t_i\Sigma d_u)^\top x_i - \alpha\sqrt{x_i^\top\left(\Sigma^{-1}+uu^\top/\sigma_w^2\right)^{-1}x_i}\right] - v_\alpha(\bar{c},\Sigma)\,. \tag{29}$$

In $\max_u \widehat{\mathbb{K}}_\alpha^N(u,\bar{c},\Sigma)$ the term $-v_\alpha(\bar{c},\Sigma)$ is constant with u, so one can omit it.

7.1. The case $N=1$. We first study the maximization of $\widehat{\mathbb{K}}_\alpha^N(\cdot,\bar{c},\Sigma)$ with $N=1$, where $\mathcal{N}(0,1)$ is reduced to a single mass point. In that case, $t_1=0$ and $p_1=1$ in (29), and we obtain the problem

$$\max_{u:||u||\le 1, x:Ax=b, x\succeq 0} \bar{c}^\top x - \alpha\sqrt{x^\top(\Sigma^{-1}+uu^\top/\sigma_w^2)^{-1}x}\ . \tag{30}$$

To get some insights on the nature of (30), suppose momentarily that we are given an optimal solution x for (30), say $\bar{x}$. Then a corresponding optimal u is given by

$$\bar{u} \in \arg\max_{u:\,||u||=1} \bar{c}^\top\bar{x} - \alpha\sqrt{\bar{x}^\top(\Sigma - \frac{\Sigma uu^\top\Sigma}{u^\top\Sigma u+\sigma_w^2})\bar{x}} = \arg\max_{u:\,||u||=1} \frac{u^\top\Sigma\bar{x}\bar{x}^\top\Sigma u}{u^\top\Sigma u+\sigma_w^2}\ .$$

This is formally equivalent to the problem solved for establishing Proposition 1, so we immediately obtain $\bar{u} = P^{-1}\Sigma\bar{x}/||P^{-1}\Sigma\bar{x}||$. The maximization of $\widehat{\mathbb{K}}_\alpha^N(\cdot,\bar{c},\Sigma)$ with $N=1$ is thus closely related to the fixed-decision heuristic, except that the reference solution $x=\bar{x}$ is now optimal for the problem with the current $\bar{c}$ and the *updated* covariance matrix Σ', which depends on u.

PROPOSITION 3. *With $\alpha>0$, the problem (30) is equivalent to the following program over $x\in\mathbb{R}^n$, $s\in\mathbb{R}$, and the symmetric matrix $W\in\mathbb{R}^{n\times n}$:*

$$\begin{aligned} &\textit{maximize} && \bar{c}^\top x - \alpha\, s \\ &\textit{subject to} && Ax=b\ ,\quad x\succeq 0\ , \\ & && \begin{bmatrix} s & x^\top \\ x & s\Sigma^{-1}+W \end{bmatrix} \succeq 0\ ,\quad \operatorname{trace}(W) = s/\sigma_w^2\ ,\quad \operatorname{rank}(W)=1\ , \end{aligned}$$

where u corresponds to a normalized dominant eigenvector of W provided $\Sigma x\neq 0$.

Proof. The constraint $\text{rank}(W) = 1$ implies that $W = \lambda uu^\top$ for some $\lambda \in \mathbb{R}$, with u corresponding to the unique normalized eigenvector of W. Since $\text{trace}(W) = \lambda$, the condition $\text{trace}(W) = s/\sigma_w^2$ implies $\lambda = s/\sigma_w^2$ and thus $W = suu^\top/\sigma_w^2$. By substitution into the SDP constraint, we have

$$\begin{bmatrix} s & x^\top \\ x & s(\Sigma^{-1} + uu^\top/\sigma_w^2) \end{bmatrix} \succeq 0 \ .$$

By the Schur complement formula, this constraint means that either $s = 0$ (and thus $x = 0$), or $s > 0$ and $s - x^\top(s[\Sigma^{-1} + uu^\top/\sigma_w^2])^{-1}x \geq 0$, that is, $s \geq \sqrt{x^\top(\Sigma^{-1} + uu^\top/\sigma_w^2)^{-1}x}$. The objective with $\alpha > 0$ ensures that s is made small, so at optimality we get $s = \sqrt{x^\top(\Sigma^{-1} + uu^\top/\sigma_w^2)^{-1}x}$. □

Proposition 3 suggests the use of a classical convexification technique where the rank-one constraint is relaxed [53], and then a solution u with $||u|| = 1$ is recovered by extracting the dominant eigenvector of W. When the rank-one constraint is relaxed, we must add the constraint $W \succeq 0$ which is no longer implied by the other constraints. Hence a first approximate solution scheme:

1. Solve the semidefinite program

$$\begin{aligned} &\text{maximize} \quad \bar{c}^\top x - \alpha\, s \qquad (31)\\ &\text{subject to} \quad Ax = b \ , \quad x \succeq 0 \ , \quad \begin{bmatrix} s & x^\top \\ x & s\Sigma^{-1} + W \end{bmatrix} \succeq 0 \ , \quad \text{trace}(W) = s/\sigma_w^2 \ , \quad W \succeq 0 \ . \end{aligned}$$

2. Return for u the normalized dominant eigenvector of W.

Step 2 is justified by the fact that the best rank-one approximation to W (in the Frobenius norm metric) is the matrix $X = \lambda_{\max}(W)uu^\top$. If W has rank one, then $\lambda_{\max}(W) = \text{trace}(W) = s/\sigma_w^2$.

In general, already with a single linear constraint $a^\top x = b$, a semidefinite programming relaxation can be arbitrarily bad [37, §13.2.4]. In our case, we have the following result.

PROPOSITION 4. *The relaxation* (31) *is tight.*

To establish the result, we first prove the following lemma.

LEMMA 6. *For an arbitrary nonzero $\bar{x} \in \mathbb{R}^n$ and for any $\Sigma^{-1}, G \succ 0$, the minimum of the semidefinite program*

$$\begin{aligned} &\textit{minimize} \quad \bar{x}^\top(\Sigma^{-1} + U)^{-1}\bar{x} \\ &\textit{subject to} \quad \text{trace}(GU) = 1,\ U \succeq 0 \end{aligned}$$

is attained by the rank-one solution

$$U^* = \frac{G^{-1}(G^{-1} + \Sigma^{-1})^{-1}xx^\top(G^{-1} + \Sigma^{-1})^{-1}G^{-1}}{x^\top(G^{-1} + \Sigma^{-1})^{-1}G^{-1}(G^{-1} + \Sigma^{-1})^{-1}x}.$$

Proof of Lemma 6. Consider first the case $G = \text{I}$. Recall that the boundary of the spectrahedron $\Omega_1 = \{U \in \mathbb{S}^n : \text{trace}(U) = 1,\ U \succeq 0\}$ is the set of rank-one matrices $\{uu^\top : ||u|| = 1\}$. Now, representing the objective with a Legendre transform, we have

$$\min_{U \in \Omega_1} \bar{x}^\top(\Sigma^{-1} + U)^{-1}\bar{x} = \min_{U \in \Omega_1} \max_y 2\bar{x}^\top y - y^\top(\Sigma^{-1} + U)y.$$

Note that: (i) Ω_1 is a nonempty compact convex set; (ii) Since $\Sigma^{-1} \succ 0$, we can confine y to a compact convex set without loss of generality; (iii) The objective is concave in y and convex in U (in fact linear in U). Therefore, we can write

$$\min_{U \in \Omega_1} \bar{x}^\top(\Sigma^{-1} + U)^{-1}\bar{x} = \max_y [\min_{U \in \Omega_1} \bar{x}^\top y - y^\top\Sigma^{-1}y - \text{trace}(yy^\top U)].$$

The inner objective being concave in U (in fact linear in U), its minimum is attained on the boundary of Ω_1. Without loss of optimality we can thus assume that U is of the form $U = uu^\top$ with $||u|| = 1$. Thus $-\text{trace}(yy^\top U) = -(u^\top y)^2$, which is minimized for $u^* = y/||y||$. The overall objective become

$$\min_{U\in\Omega_1} \bar{x}^\top(\Sigma^{-1}+U)^{-1}\bar{x} = \max_y 2\bar{x}^\top y - y^\top(\mathrm{I}+\Sigma^{-1})y = \bar{x}^\top(\mathrm{I}+\Sigma^{-1})^{-1}\bar{x},$$

where the maximum is attained for $y^* = (\mathrm{I}+\Sigma^{-1})^{-1}\bar{x}$. Setting $B = (\mathrm{I}+\Sigma^{-1})^{-1}$, the expression for U^* follows immediately:

$$U^* = u^*u^{*\top} = y^*y^{*\top}/(y^{*\top}y^*) = B\bar{x}\bar{x}^\top B/\bar{x}^\top B^\top B\bar{x}.$$

Consider now the case with a general $G \succ 0$. Since $\{d\in\mathbb{R}^n : \text{trace}(Gdd^\top) = 1\} = \{d' = G^{-1/2}u' : ||u'|| = 1\}$, the boundary of $\Omega_G = \{U\in\mathbb{S}^n : \text{trace}(GU) = 1,\ U\succeq 0\}$ is the set of rank-one matrices $G^{-1/2}uu^\top G^{-1/2}$ with $||u|| = 1$. This leads to the representation $\Omega_G = \{U = G^{-1/2}VG^{-1/2} : V\in\Omega_1\}$. Therefore, the minimization problem over $U\in\Omega_G$ can be recast as

$$\min_{V\in\Omega_1} \bar{x}^\top(\Sigma^{-1}+G^{-1/2}VG^{-1/2})^{-1}\bar{x} = \min_{V\in\Omega_1} \bar{x}^\top G^{1/2}([G^{1/2}\Sigma^{-1}G^{1/2}+V])^{-1}G^{1/2}\bar{x}.$$

Using the substitution $\bar{x}\mapsto G^{1/2}\bar{x}$, and $\Sigma^{-1}\mapsto G^{1/2}\Sigma^{-1}G^{1/2}$ in the case for $G = \mathrm{I}$, we obtain the optimal solution $V^* = B\bar{x}\bar{x}^\top B/\bar{x}^\top B^\top B\bar{x}$ with $B := (\mathrm{I}+G^{1/2}\Sigma^{-1}G^{1/2})^{-1}G^{1/2} = G^{-1/2}(G^{-1}+\Sigma^{-1})^{-1}$ and thus $U^* = G^{-1/2}V^*G^{-1/2}$. □

Proof of Proposition 4. If we set $W = sU/\sigma_w^2$, assuming that $s > 0$, the problem (31) becomes

$$\text{maximize}\quad \bar{c}^\top x - \alpha s \quad \text{subject to}\quad Ax = b,\ x\succeq 0,\ s^2 \geq x^\top(\Sigma^{-1}+U)^{-1}x,\ \text{trace}(U) = 1,\ U\succeq 0.$$

For any fixed x, the best value of the objective is obtained by minimizing s, which in turns leads to the minimization of $f_x(U) = x^\top(\Sigma^{-1}+U)^{-1}x$ subject to $\text{trace}(U) = 1$ and $U\succeq 0$. By Lemma 6, the minimum of $f_x(U)$ is attained by a rank-one matrix U, so there also exists an optimal rank-one matrix $W = sU/\sigma_w^2$. When $s = 0$, we have $W = 0$ from $\text{trace}(W) = 0/\sigma_w^2$ and $W\succeq 0$. □

We have thus established that there exists an optimal solution to (31) where W has rank one. We also observe a preference for the rank-one solution. To see this, let $\nu\in\mathbb{R}^n$ with elements $\nu_1\geq\cdots\geq\nu_n\geq 0$ denote the vector of sorted eigenvalues of W. We have $\text{trace}(W) = \sum_{i=1}^n \nu_i = \sum_{i=1}^n |\nu_i| = ||\nu||_1$. Since L1-norm regularization induces sparsity in the solution, one can see that the constraint $\text{trace}(W) = s/\sigma_w^2$, combined with the fact that s is minimized in the objective, has a beneficial effect on the formulation: it induces zero eigenvalues in W, and thus rank reduction. Nuclear norm minimization, or trace minimization in the special case of positive semidefinite matrices, is a convex technique for inducing low-rank solutions [21]; in our case the trace minimization effect is a byproduct of the original objective.

Another solution approach to (30) is also possible, which directly exploits the structure of the optimal solution for u.

PROPOSITION 5. *Define C such that $C^\top C = P^{-1}\Sigma$, where as usual $P = (\sigma_w^2\mathrm{I}_n+\Sigma)$. Then, an optimal solution (x^*,u^*) to the problem (30) can be obtained by solving the following conic program over $(x,s)\in\mathbb{R}^n\times\mathbb{R}$,*

$$\begin{aligned} &\textit{maximize} && \bar{c}^\top x - \alpha s \\ &\textit{subject to} && Ax = b\ ,\quad x\succeq 0\ , \\ &&& ||Cx|| \leq \sigma_w^{-1}s, \end{aligned} \tag{32}$$

then setting $u^* = P^{-1}\Sigma x^*/||P^{-1}\Sigma x^*||$.

Proof. As explained at the beginning of this section, the vector $\bar{u} = P^{-1}\Sigma\bar{x}/||P^{-1}\Sigma\bar{x}||$ is optimal given $\bar{x}$. This means that at optimality,

$$\bar{s}^2 = \bar{x}^\top \left[\Sigma^{-1} + \bar{u}\bar{u}^\top/\sigma_w^2\right]^{-1} \bar{x} = \bar{x}^\top \left[\Sigma - \frac{\Sigma P^{-1}\Sigma\bar{x}\bar{x}^\top\Sigma P^{-1}\Sigma}{\bar{x}^\top\Sigma P^{-1}PP^{-1}\Sigma\bar{x}}\right]\bar{x}$$

$$= \bar{x}^\top\Sigma\bar{x} - \frac{(\bar{x}^\top\Sigma P^{-1}\Sigma\bar{x})^2}{\bar{x}^\top\Sigma P^{-1}\Sigma\bar{x}} = \bar{x}^\top(\Sigma - \Sigma P^{-1}\Sigma)\bar{x} = \bar{x}^\top(\sigma_w^2 P^{-1}\Sigma)\bar{x} = \sigma_w^2||C\bar{x}||^2 \ ,$$

where we have defined C such that $C^\top C := P^{-1}\Sigma$. The subproblem for optimizing x follows immediately. □

7.2. The case $N > 1$, $\alpha = 0$. When $N > 1$, the problem takes into account the update of $\bar{c}$ to $\bar{c}'$, which depends on t and u. The following lemma is instrumental for dealing with the nonlinear dependence of d_u on u, as defined in (22). From Theorem 1, we know we can restrict our attention to measurements u with $||u|| = 1$.

LEMMA 7. *The nonconvex set*

$$D = \left\{ d = \frac{u}{\sqrt{u^\top\Sigma u + \sigma_w^2}} : ||u|| = 1, u \in \mathbb{R}^n \right\} \tag{33}$$

admits the alternative representations

$$D = \{d' = P^{-1/2}u' : ||u'|| = 1, u' \in \mathbb{R}^n\}, \tag{34}$$

$$D = \{d'' \in \mathbb{R}^n : \text{trace}(Pd''d''^\top) = 1\}. \tag{35}$$

Proof. If $d \in D$, there exists $u \in \mathbb{R}^n$ with $u^\top u = 1$ such that

$$d = \frac{u}{\sqrt{u^\top\Sigma u + \sigma_w^2}} = \frac{u}{\sqrt{u^\top Pu}} = \frac{P^{-1/2}P^{1/2}u}{||P^{1/2}u||} = P^{-1/2}u'$$

where $u' = P^{1/2}u/||P^{1/2}u||$ satisfies $||u'|| = 1$, showing (33) → (34). Conversely, if $d' \in D$, there exists $u' \in \mathbb{R}^n$ with $u'^\top u' = 1$ such that

$$d' = P^{-1/2}u' = ||P^{-1/2}u'||v$$

where we have defined $v = P^{-1/2}u'/||P^{-1/2}u'||$; then noting that $v^\top v = 1$, we evaluate

$$[v^\top\Sigma v + \sigma_w^2]^{-1/2} = [v^\top Pv]^{-1/2} = \left[\frac{u'^\top P^{-1/2}PP^{-1/2}u'}{||P^{-1/2}u'||^2}\right]^{-1/2} = ||P^{-1/2}u'||,$$

so that $d' = [v^\top\Sigma v + \sigma_w^2]^{-1/2}v$, showing (34) → (33) with $u = v = P^{-1/2}u'/||P^{-1/2}u'||$. This establishes the equivalence between (33) and (34).

The well-know identity $\{Q^{1/2}z : ||z|| = 1, z \in \mathbb{R}^n\} = \{z \in \mathbb{R}^n : z^\top Q^{-1}z = 1\}$ applied to $Q = P^{-1}$, and the relation $z^\top Q^{-1}z = \text{trace}(z^\top Q^{-1}z) = \text{trace}(Q^{-1}zz^\top) = \text{trace}(Pzz^\top)$, establish the equivalence between (34) and (35). □

The following lemma, due to [58], will be useful to strengthen the relaxations.

LEMMA 8. *Assume $\mathcal{X} = \{x \in \mathbb{R}^n : Ax = b, x \succeq 0\}$ is bounded and not reduced to $\{0\}$. Fix $\nu \in \mathbb{R}^n$ with $\nu_i > 0$ for each i, and define $\bar{\gamma}_\nu = \sup_{x\in\mathcal{X}} \nu^\top x$. Then the following relation holds true for any $x \in \mathcal{X}$:*

$$xx^\top \preceq \bar{\gamma}_\nu \text{Diag}(x)\text{Diag}(\nu)^{-1},$$

where $\text{Diag}(z)$ *denotes the diagonal matrix with elements z_i.*

Proof. Since $x \succeq 0$ and $\mathcal{X} \neq \{0\}$, $\bar{\gamma}_\nu > 0$. Since $\mathcal{X}$ is bounded, $\bar{\gamma}_\nu < \infty$. A lemma established in [58] shows that for any $x \in \mathcal{X}$, $\mathrm{diag}(\nu) x x^\top \mathrm{diag}(\nu) \preceq \bar{\gamma}_\nu \mathrm{diag}(\nu)\mathrm{diag}(x)$. Recall that $S \succeq 0$ iff $PSP^\top \succeq 0$, where P can be any invertible matrix. Applying this rule to the inequality with $P = \mathrm{diag}\{\nu\}^{-1}$ establishes the result. $\square$

We have now the necessary ingredients for proposing a solution scheme to (11), first in the case $\alpha = 0$. As usual, $P = \Sigma + \sigma_w^2 \mathrm{I}_n$.

1. Choose a quantization $\{p_i, t_i\}_{i=1}^N$ of $t \sim \mathcal{N}(0,1)$.

Construct the symmetric matrices

$$C_i = \frac{1}{2}\left[\begin{array}{c|c|c} 0 & \bar{c}^\top & 0^\top \\ \hline \bar{c} & 0 & t_i\Sigma \\ \hline 0 & t_i\Sigma & 0 \end{array}\right] \in \mathbb{R}^{(2n+1)\times(2n+1)}, \qquad 1 \leq i \leq N.$$

2. Generate a set of vectors $\{\nu_\ell\}_{\ell=1}^M$, $\nu_\ell \succ 0$, and evaluate

$$\bar{\gamma}_\ell = \max_{x \in \mathcal{X}} \nu_\ell^\top x.$$

3. Solve the following SDP over the symmetric optimization matrices $Y \in \mathbb{R}^{n\times n}$ and

$$Z_i = \left[\begin{array}{c|c|c} Z_i^{11} & Z_i^{1x} & Z_i^{1d} \\ \hline Z_i^{x1} & Z_i^{xx} & Z_i^{xd} \\ \hline Z_i^{d1} & Z_i^{dx} & Z_i^{dd} \end{array}\right] = \left[\begin{array}{c|c|c} 1 & x_i^\top & d^\top \\ \hline x_i & Z_i^{xx} & Z_i^{xd} \\ \hline d & Z_i^{dx} & Y \end{array}\right] \in \mathbb{R}^{(2n+1)\times(2n+1)}, \qquad 1 \leq i \leq N \ :$$

$$\begin{aligned} \text{maximize} \quad & \textstyle\sum_{i=1}^N p_i \mathrm{trace}(C_i Z_i) \\ \text{subject to} \quad \forall\, i: \quad & Z_i \succeq 0, \\ & Z_i^{11} = 1, \quad A Z_i^{x1} = b, \quad Z_i^{x1} \succeq 0, \\ & A Z_i^{xx} A^\top = b b^\top, \ [Z_i^{xx}]_{qr} \geq 0 \quad \forall q, r, \\ & Z_i^{xx} \preceq \bar{\gamma}_\ell \mathrm{Diag}(Z_i^{x1})\mathrm{Diag}(\nu_\ell)^{-1} \quad \forall \ell, \\ & Z_i^{dd} = Y, \\ & \mathrm{trace}(PY) = 1. \end{aligned}$$

4. Return for u the eigenvector associated to the largest eigenvalue of Y.

The scheme is based on the relation

$$(\bar{c} + t_i \Sigma d)^\top x_i = \frac{1}{2}\mathrm{trace}\left(\begin{bmatrix} 0 & \bar{c}^\top & 0^\top \\ \bar{c} & 0 & t_i\Sigma \\ 0 & t_i\Sigma & 0 \end{bmatrix}\begin{bmatrix} 1 \\ x_i \\ d \end{bmatrix}\begin{bmatrix} 1 \\ x_i \\ d \end{bmatrix}^\top\right) \ ,$$

where we define

$$Z_i = \begin{bmatrix} 1 \\ x_i \\ d \end{bmatrix}\begin{bmatrix} 1 \\ x_i \\ d \end{bmatrix}^\top = \begin{bmatrix} 1 & x_i^\top & \dfrac{u^\top}{\sqrt{u^\top\Sigma u + \sigma_w^2}} \\ x_i & x_i x_i^\top & \dfrac{x_i u^\top}{\sqrt{u^\top\Sigma u + \sigma_w^2}} \\ \dfrac{u}{\sqrt{u^\top\Sigma u + \sigma_w^2}} & \dfrac{u x_i^\top}{\sqrt{u^\top\Sigma u + \sigma_w^2}} & \dfrac{u u^\top}{u^\top\Sigma u + \sigma_w^2} \end{bmatrix},$$

which is semidefinite positive and has rank 1.

The constraints $Ax = b$ and $x \succeq 0$ imply $Ax_i x_i^\top A^\top = bb^\top$ (linear equality between matrices) and $[x_i x_i^\top]_{qr} \geq 0$ for $1 \leq q, r \leq n$ (nonnegativity of the matrix $x_i x_i^\top$). In terms of the matrix Z_i, we write $AZ_i^{xx}A^\top = bb^\top$ and $[Z_i^{xx}]_{qr} \geq 0$. The constraint $\text{trace}(Z_i^{xx}AA^\top) = b^\top b$ would be of no use here because it is implied by $AZ_i^{xx}A^\top = bb^\top$, so we use Lemma 8 to further control Z_i^{xx} by the constraint $Z_i^{xx} \preceq \bar{\gamma}_\ell \text{Diag}(Z_i^{x1})\text{Diag}(\nu_\ell)^{-1}$. A single inequality suffices since we impose $Z_i^{x1} \in \mathcal{X}$, which is bounded by assumption. Introducing additional valid inequalities can strengthen the relaxation but can also increase the rank of the solution Y, since the minimal rank solution is affected by the number of constraints [40, 4].

We introduce the variable $Y = uu^\top/(u^\top \Sigma u + \sigma_w^2)$ to write the constraints $Z_i^{uu} = Y = Z_j^{uu}$, $1 \leq i, j \leq N$. From Theorem 1 we want $||u|| = 1$. From Lemma 7, this is possible by imposing $\text{trace}(PY) = 1$ and $\text{rank}(Y) = 1$. All the rank-one constraints are then relaxed. We obtain our approximation of the optimal u through the normalized eigenvector associated to the largest eigenvalue of Y, since we have $Yu = (u^\top u/u^\top Pu)u = \lambda u$ with $\lambda = u^\top Pu$ when Y follows its rank-one definition.

When the feasible set $\mathcal{X}$ can be expressed in terms of the squares of the coordinates of x, then this representation should be used to create constraints on Z_i^{xx}. For example, box constraints $\{0 \preceq [x]_j \preceq [b]_j\}$ imply $\{0 \preceq [x]_j^2 \preceq [b]_j^2\}$ and thus $\text{diag}(Z_i^{xx}) \leq \text{diag}(b)^2$. Binary constraints with value ± 1 would imply $\text{diag}(Z_i^{xx}) = 1$.

It is also insightful to state the problem as a general nonconvex quadratic program with quadratic constraints (QCQP). Recalling Corollary 1 and discretizing the expectation, one would obtain

$$\begin{aligned} \text{maximize} \quad & \bar{c}^\top \left(\textstyle\sum_{i=1}^N p_i x_i\right) + t \\ \text{subject to} \quad & z = \textstyle\sum_{i=1}^N p_i t_i x_i \\ & t^2 - z^\top Q z \leq 0, \quad t \geq 0, \\ & Ax_i = b, \quad x \succeq 0, \quad 1 \leq i \leq N, \end{aligned}$$

where $Q = \Sigma P^{-1} \Sigma \succeq 0$ makes the quadratic constraint utterly nonconvex. The constraint $t \geq 0$ is redundant given the objective function. A corresponding optimal u^* is given by $u^* = P^{-1}\Sigma z^*/||P^{-1}\Sigma z^*||$.

7.3. General case: $N > 1$, $\alpha > 0$. The solution scheme for the general case combines the techniques used in the two preceding cases.

1. Choose a quantization $\{p_i, t_i\}_{i=1}^N$ of $t \sim \mathcal{N}(0,1)$.

Define the symmetric matrices

$$C_i = \frac{1}{2}\begin{bmatrix} 0 & \bar{c}^\top & 0^\top \\ \bar{c} & 0 & t_i\Sigma \\ 0 & t_i\Sigma & 0 \end{bmatrix} \in \mathbb{R}^{(2n+1)\times(2n+1)}, \qquad 1 \leq i \leq N.$$

2. Generate a set of vectors $\{\nu_\ell\}_{\ell=1}^M$, $\nu_\ell \succ 0$, and evaluate $\bar{\gamma}_\ell = \max_{x\in\mathcal{X}} \nu_\ell^\top x$.
3. Solve the following SDP over $u \in \mathbb{R}^n$, $s_i \in \mathbb{R}$ and the symmetric matrices Y, $W_i \in \mathbb{R}^{n\times n}$, and

$$Z_i = \begin{bmatrix} Z_i^{11} & Z_i^{1x} & Z_i^{1d} \\ Z_i^{x1} & Z_i^{xx} & Z_i^{xd} \\ Z_i^{d1} & Z_i^{dx} & Z_i^{dd} \end{bmatrix} = \begin{bmatrix} 1 & x_i^\top & d^\top \\ x_i & Z_i^{xx} & Z_i^{xd} \\ d & Z_i^{dx} & Y \end{bmatrix} \in \mathbb{R}^{(2n+1)\times(2n+1)} \quad , \quad 1 \leq i \leq N \quad :$$

$$\begin{aligned} \text{maximize} \quad & \textstyle\sum_{i=1}^N p_i[\text{trace}(C_i Z_i) - \alpha s_i] \\ \text{subject to} \quad & \text{trace}(PY) = 1 \quad , \\ \forall i: \quad & Z_i \succeq 0 \quad , \end{aligned}$$

$$
\begin{aligned}
&\text{trace}(W_i) = s_i/\sigma_w^2 \ , \\
&\begin{bmatrix} s_i & Z_i^{1x} \\ Z_i^{x1} & s_i\Sigma^{-1} + W_i \end{bmatrix} \succeq 0 \ , \\
&\begin{bmatrix} W_i & w_i \\ w_i^\top & 1 \end{bmatrix} \succeq 0 \ , \\
&\begin{bmatrix} Y & w_i \\ w_i^\top & \text{trace}(PW_i) \end{bmatrix} \succeq 0 \ , \\
&Z_i^{11} = 1 \ , \quad AZ_i^{x1} = b \ , \quad Z_i^{x1} \succeq 0 \ , \\
&AZ_i^{xx}A^\top = bb^\top \ , \quad [Z_i^{xx}]_{qr} \geq 0 \quad \forall q, r \ , \\
&Z_i^{xx} \preceq \bar{\gamma}_\ell \text{Diag}(Z_i^{x1})\text{Diag}(\nu_\ell)^{-1} \quad \forall \ell \ , \\
&Z_i^{dd} = Y \ .
\end{aligned}
$$

4. Return for u the eigenvector associated to the largest eigenvalue of Y.

In the SDP, using $||u|| = 1$ we define $Y = dd^\top = uu^\top/u^\top Pu = uu^\top/\text{trace}(Puu^\top)$. We have $\text{trace}(PY) = \text{trace}(u^\top Pu/u^\top Pu) = 1$. For each i, we define $s_i \geq 0$ and $w_i w_i^\top = W_i = s_i uu^\top/\sigma_w^2$. We have $\text{trace}(W_i) = s_i/\sigma_w^2$. Assuming $s_i > 0$, we have $uu^\top = \sigma_w^2 W_i/s_i$, so we can rewrite Y as

$$
Y = \frac{\sigma_w^2 W_i/s_i}{\text{trace}(P\sigma_w^2 W_i/s_i)} = \frac{w_i w_i^\top}{\text{trace}(PW_i)} \ .
$$

We relax the definitions of W_i and Y to $W_i \succeq w_i w_i^\top$ and $Y \succeq w_i w_i^\top/\text{trace}(PW_i)$, which can be expressed, using a Schur complement logic, by the constraints

$$
\begin{bmatrix} W_i & w_i \\ w_i^\top & 1 \end{bmatrix} \succeq 0 \ , \quad \begin{bmatrix} Y & w_i \\ w_i^\top & \text{trace}(PW_i) \end{bmatrix} \succeq 0 \ .
$$

The rest of the construction of the program follows the logic of Sections 7.1 and 7.2.

8. Convergence. In this section, we show a form of asymptotic convergence for the expected improvement policy. The main result is that the quantity $\mathbb{K}_\alpha(u, \bar{c}_k, \Sigma_k)$ converges to zero for all points u on the L2 sphere. We also show that, as a consequence, we have $x^\top \Sigma_k x \to 0$ for all $x \in \mathcal{X}$, which means that the objective value $x^\top c^{\text{true}}$ of every feasible implementation decision x is learned perfectly (with zero variance) in the limit. Unlike the policy studied in Section 6, this does not necessarily mean that the posterior covariance Σ_k converges to zero under the KG policy; rather, it means that we obtain perfect information about every decision of interest.

We assume that $\mathcal{X}$ is bounded and the risk-aversion parameter $\alpha > 0$. We also require one simplifying technical assumption for Propositions 9 and 10: we assume that our prior covariance matrix is given by $\Sigma_0 = \beta \text{I}_n$ for some constant $\beta > 0$. A consequence of this assumption is that $u^T \Sigma_0 u = \beta$ for all u on the L2 sphere. The assumption can be a reasonable choice for some applications of recursive least squares; for example, [41] recommends using this initialization in MDPs with basis function approximations.

We begin by showing that the expected improvement in the direction u is bounded above by a function of the maximum variance reduction possible by measuring u.

PROPOSITION 6. *For any u,*

$$
\mathbb{K}_\alpha(u, \bar{c}, \Sigma) \leq \left(\alpha + \frac{2}{\sqrt{2\pi}}\right) \max_{x \in \mathcal{X}} \left| x^\top \Sigma d_u \right| .
$$

Proof. We write

$$\begin{aligned}\mathbb{K}_\alpha(u,\bar{c},\Sigma) &= \mathbb{E}\left\{\max_{x\in\mathcal{X}} \bar{c}^\top x - \alpha\sqrt{x^\top\Sigma' x} + tx^\top\Sigma d_u\right\} - v_\alpha(\bar{c},\Sigma)\\ &\leq v_\alpha(\bar{c},\Sigma') - v_\alpha(\bar{c},\Sigma) + \mathbb{E}\left\{\max_{x\in\mathcal{X}} tx^\top\Sigma d_u\right\}.\end{aligned}$$

Now, observe that

$$\begin{aligned}v_\alpha(\bar{c},\Sigma') - v_\alpha(\bar{c},\Sigma) &\leq \max_{x\in\mathcal{X}} \alpha\left(\sqrt{x^\top\Sigma x} - \sqrt{x^\top\Sigma' x}\right)\\ &\leq \max_{x\in\mathcal{X}} \alpha\sqrt{x^\top(\Sigma-\Sigma')x}\\ &= \max_{x\in\mathcal{X}} \alpha\left|x^\top\Sigma d_u\right|.\end{aligned}$$

Furthermore,

$$\begin{aligned}\mathbb{E}\left\{\max_{x\in\mathcal{X}} tx^\top\Sigma d_u\right\} &= \mathbb{E}\left\{\max_{x\in\mathcal{X}} t1_{\{t\geq 0\}}x^\top\Sigma d_u\right\} + \mathbb{E}\left\{\max_{x\in\mathcal{X}} t1_{\{t<0\}}x^\top\Sigma d_u\right\}\\ &= \left(\max_{x\in\mathcal{X}} x^\top\Sigma d_u\right)\mathbb{E}\left\{t1_{\{t\geq 0\}}\right\} + \left(\min_{x\in\mathcal{X}} x^\top\Sigma d_u\right)\mathbb{E}\left\{t1_{\{t<0\}}\right\}\\ &= \frac{1}{\sqrt{2\pi}}\left(\max_{x\in\mathcal{X}} x^\top\Sigma d_u - \min_{x\in\mathcal{X}} x^\top\Sigma d_u\right)\\ &\leq \frac{2}{\sqrt{2\pi}}\max_{x\in\mathcal{X}}\left|x^\top\Sigma d_u\right|,\end{aligned}$$

which completes the proof. □

By the Cauchy-Schwarz inequality, it follows that, if $\{\bar{c}_k,\Sigma_k\}$ is a sequence satisfying $u^\top\Sigma_k u\to 0$, we also have $\mathbb{K}_\alpha(u,\bar{c}_k,\Sigma_k)\to 0$. If the variance of our beliefs about u decreases to zero (for example, if we measure u infinitely often), the expected improvement in this direction also vanishes. Our next result is related to the converse of this statement.

PROPOSITION 7. *Let u be a point on the L2 sphere with $\mathbb{K}_\alpha(u,\bar{c},\Sigma)=0$. Then, $x^\top\Sigma u=0$ for all $x\in\mathcal{X}$.*

Proof. The function $h(t)=v_\alpha(\bar{c}+t\Sigma d_u,\Sigma')$ is a maximum of linear functions of t, and therefore is convex. Thus, by Jensen's inequality,

$$\mathbb{E}v_\alpha(\bar{c}+t\Sigma d_u,\Sigma')\geq v_\alpha(\bar{c},\Sigma').$$

Letting $\bar{x}=\arg\max_{x\in\mathcal{X}}\bar{c}^\top x-\alpha\sqrt{x^\top\Sigma x}$, we have

$$\begin{aligned}v_\alpha(\bar{c},\Sigma') &\geq \bar{c}^\top\bar{x} - \alpha\sqrt{\bar{x}^\top\Sigma'\bar{x}}\\ &\geq v_\alpha(\bar{c},\Sigma),\end{aligned}$$

since the variance of our beliefs about any x is always decreasing after each measurement.

However, since we assume that $\mathbb{K}_\alpha(u,\bar{c},\Sigma)=0$, all of the above inequalities must hold with equality. Consequently, it follows that $\bar{x}^\top\Sigma u=0$. For this reason,

$$\bar{c}^\top\bar{x} - \alpha\sqrt{\bar{x}^\top\Sigma'\bar{x}} + t\bar{x}^\top\Sigma d_u = \bar{c}^\top\bar{x} - \alpha\sqrt{\bar{x}^\top\Sigma\bar{x}} = v_\alpha(\bar{c},\Sigma)$$

almost surely. We can then write

$$v_\alpha(\bar{c}+t\Sigma d_u,\Sigma') = \max\left\{v_\alpha(\bar{c}+t\Sigma d_u,\Sigma'),v_\alpha(\bar{c},\Sigma)\right\}.$$

Since the expected improvement is zero, it follows that

$$\mathbb{E}\left\{\max\left\{v_\alpha(\bar{c}+t\Sigma d_u,\Sigma'),v_\alpha(\bar{c},\Sigma)\right\}-v_\alpha(\bar{c},\Sigma)\right\}=0.$$

However, the random variable inside the expectation is a.s. positive, whence

$$\max\left\{v_\alpha(\bar{c}+t\Sigma d_u,\Sigma'),v_\alpha(\bar{c},\Sigma)\right\}=v_\alpha(\bar{c},\Sigma)$$

and

$$\bar{c}^\top x-\alpha\sqrt{x^\top\Sigma' x}+tx^\top\Sigma d_u\leq v_\alpha(\bar{c},\Sigma),$$

almost surely, for all $x\in\mathcal{X}$. However, since t can take on any real value, this is only possible if $x^\top\Sigma u=0$ for every $x\in\mathcal{X}$. □

From Propositions 6 and 7, it follows that $\mathbb{K}_\alpha(u,\bar{c},\Sigma)=0$ if and only if $u^\top\Sigma x=0$ for all $x\in\mathcal{X}$. In particular, if $x\in\mathcal{X}$, we have zero expected improvement along $\frac{x}{\|x\|}$ if and only if $x^\top\Sigma x=0$. Our next result connects this limiting case to the asymptotic behavior of the expected improvement.

PROPOSITION 8. *For fixed u, the expected improvement $\mathbb{K}_\alpha(u,\bar{c},\Sigma)$ is continuous in $\bar{c}$ and Σ.*

Proof. We first address continuity in $\bar{c}$. Observe that, for any fixed t, $v_\alpha(\bar{c}+t\Sigma d_u,\Sigma')$ is convex in $\bar{c}$, because it is a maximum of linear (and thus convex) functions of $\bar{c}$. Taking expectations preserves convexity, so $\mathbb{K}_\alpha(u,\bar{c},\Sigma)$ is convex in $\bar{c}$. However, convex functions are continuous in the interior of their domain, which in the case of $\bar{c}$ is all of $\mathbb{R}^n$.

Next, we address continuity in Σ. Let $\{\Sigma_k\}$ be a sequence of positive semidefinite matrices that converges componentwise to a positive semidefinite matrix Σ_∞. Let $t\sim\mathcal{N}(0,1)$ and define random variables $T_k=v_\alpha(u,\bar{c}+t\Sigma_k d_u,\Sigma'_k)$ for $k=1,2,...$ and $k=\infty$. We show that $T_k\to T_\infty$ in L^1 by writing

$$\begin{aligned}\mathbb{E}\left|T_k-T_\infty\right|&\leq\mathbb{E}\max_{x\in\mathcal{X}}\left|\alpha\left(\sqrt{x^\top\Sigma'_\infty x}-\sqrt{x^\top\Sigma'_k x}\right)+t\left(\frac{x^\top\Sigma_k u}{\sqrt{\sigma_w^2+u^\top\Sigma_k u}}-\frac{x^\top\Sigma_\infty u}{\sqrt{\sigma_w^2+u^\top\Sigma_\infty u}}\right)\right|\\&\leq\alpha\max_{x\in\mathcal{X}}\sqrt{x^\top\left(\Sigma'_\infty-\Sigma'_k\right)x}+\left(\mathbb{E}\left|t\right|\right)\max_{x\in\mathcal{X}}\left|\frac{x^\top\Sigma_k u}{\sqrt{\sigma_w^2+u^\top\Sigma_k u}}-\frac{x^\top\Sigma_\infty u}{\sqrt{\sigma_w^2+u^\top\Sigma_\infty u}}\right|.\end{aligned}$$

For any $\varepsilon>0$ and large enough k, we will have $\mathbb{E}\left|T_k-T_\infty\right|\leq(\alpha+\mathbb{E}\left|t\right|)\varepsilon$, completing the proof. □

This result has two consequences. First, $\mathbb{K}_\alpha(u,\bar{c}_k,\Sigma_k)$ always has a limit for any u. This occurs because, for any sequence of measurements, we can write $\bar{c}_k=\mathbb{E}\left(c^{\text{true}}\,|\,\mathcal{F}_k\right)$ and $\Sigma_k=\mathbb{E}\left(c^{\text{true}}\left(c^{\text{true}}\right)^\top|\,\mathcal{F}_k\right)$, where $\mathcal{F}_k$ is the sigma-algebra generated by the first k measurements and their outcomes. Therefore, by martingale convergence theory, both $\bar{c}_k$ and Σ_k have a.s. limits.

The second main consequence is that $\mathbb{K}_\alpha(u,\bar{c}_k,\Sigma_k)\to 0$ if and only if $u^\top\Sigma x\to 0$ for all $x\in\mathcal{X}$. This follows from Propositions 6, 7, and 8.

Our next objective is to show that the posterior variance of our beliefs converges to zero for vectors that are accumulation points of the sequence of measurements. This is done in the following two propositions, both of which rely on the assumption that $\Sigma_0=\beta\mathrm{I}_n$.

PROPOSITION 9. *Given some fixed $\bar{u}$ on the L2 sphere and some small $\varepsilon>0$, let $B=\{u:\|u\|=1,\ \|u-\bar{u}\|<\varepsilon\}$. Consider an arbitrary $u\in B$ and define*

$$\tilde{u}=\arg\min_{u'\in B}\left|u^\top\Sigma_0 u'\right|=\arg\min_{u'\in B}\left|u^\top u'\right|.$$

Note that, for small enough ε, $\tilde{u}$ cannot be orthogonal to u. Suppose that, at time k', a total of k measurements have been made in the set B. Then, it follows that the posterior variance of our beliefs about u satisfies the inequality

$$u^\top \Sigma_{k'} u \leq \beta - \frac{\beta_0^2 k}{\beta k + \sigma_w^2} \tag{36}$$

where $\beta_0 = \tilde{u}^\top \Sigma_0 u$.

Proof. The posterior variance $u^\top \Sigma_k u$ is monotonically decreasing in k, and depends only on the vectors we measure, not on the observations. Because any measurement decreases the variance, the posterior variance at time k is bounded above by the matrix created by applying (8) only after those measurements u_k that are in the set B. All measurements outside B can be ignored, because they only decrease the variance further.

Consider a policy that measures $u_1 = u_2 = ... = u_k = \tilde{u}$. Using the Sherman-Morrison formula, we find that

$$u^\top \Sigma_k u = \beta - \frac{\beta_0^2 k}{\beta k + \sigma_w^2}.$$

Suppose that $u_{k+1} = \tilde{u}$ as well. Then, the variance reduction in our beliefs about u, achieved between time k and time $k+1$, is given by

$$u^\top \Sigma_k u - u^\top \Sigma_{k+1} u = \frac{\beta_0^2 \sigma_w^2}{\left((k+1)\beta + \sigma_w^2\right)\left(k\beta + \sigma_w^2\right)}.$$

Now consider a situation where $u_{k+1} = u'$ for some $u' \in B$. In this case, it can be worked out that

$$u^\top \Sigma_k u - u^\top \Sigma_{k+1} u = \left(\beta_2 - \frac{k\beta_0\beta_1}{k\beta + \sigma_w^2}\right)^2 \left(\sigma_w^2 + \beta - \frac{k\beta_1^2}{k\beta + \sigma_w^2}\right)^{-1}, \tag{37}$$

where $\beta_1 = \tilde{u}^\top \Sigma_0 u'$ and $\beta_2 = u^\top \Sigma_0 u'$.

We now study the numerator of (37). Observe that

$$\left(\beta_2 - \frac{k\beta_0\beta_1}{k\beta + \sigma_w^2}\right)^2 = \left|\beta_2 - \frac{k\beta_0\beta_1}{k\beta + \sigma_w^2}\right|^2$$

and

$$\left|\beta_2 - \frac{k\beta_0\beta_1}{k\beta + \sigma_w^2}\right| \geq |\beta_2| - \frac{k\,|\beta_0| \cdot |\beta_1|}{k\beta + \sigma_w^2}. \tag{38}$$

Note that the right-hand side of (38) is positive because $|\beta_0| \leq |\beta_2|$ by the definition of $\tilde{u}$, and $|\beta_1| \leq \beta$ by the Cauchy-Schwarz inequality. Consequently, (37) leads to

$$u^\top \Sigma_k u - u^\top \Sigma_{k+1} u \geq \left(|\beta_2| - \frac{k\,|\beta_0| \cdot |\beta_1|}{k\beta + \sigma_w^2}\right)^2 \left(\sigma_w^2 + \beta - \frac{k\,|\beta_1|^2}{k\beta + \sigma_w^2}\right)^{-1}.$$

Now, it is possible to apply the arguments given in the proof of Proposition 5.3 in [52], which show that the variance reduction obtained when $u_{k+1} = u'$ is greater than the variance reduction obtained when $u_{k+1} = \tilde{u}$. This is true for all k. Consequently, the smallest variance reduction that is possible with k measurements in the set B is achieved by always measuring $\tilde{u}$. The bound in (36) follows. □

Now suppose that the reference point $\bar{u}$ is an accumulation point of the sequence $\{u_k\}$ of measurements. We know that such a point must exist since the sequence is bounded. In this case, we know that infinitely many measurements will be made inside the set $B = \{u : \|u\| = 1,\ \|u - \bar{u}\| < \varepsilon\}$. Letting Σ_∞ be the limit of the sequence (Σ_k) of posterior covariance matrices, and applying Proposition 9 to the point $\bar{u}$, we find that

$$\bar{u}^\top \Sigma_\infty \bar{u} \leq \beta \left(1 - \min_{u \in B} \left(\bar{u}^\top u\right)^2\right). \tag{39}$$

This leads to the following limiting result.

PROPOSITION 10. *Let $\bar{u}$ be an accumulation point of the sequence of measurements. Then, $\bar{u}^\top \Sigma_\infty \bar{u} = 0$.*

Proof. We rewrite the set B as

$$\begin{aligned} B &= \left\{u : \|u\| = 1,\ u^\top u - 2u^\top \bar{u} + \left(\bar{u}^\top \bar{u} - \varepsilon^2\right) \leq 0\right\} \\ &= \left\{u : \|u\| = 1,\ u^\top \bar{u} \geq \left(1 - \frac{1}{2}\varepsilon^2\right)\right\}. \end{aligned}$$

For small enough ε, $\left(1 - \frac{1}{2}\varepsilon^2\right) > 0$. Then,

$$\min_{u \in B} \left(\bar{u}^\top u\right)^2 = \left(\min_{u \in B'} \bar{u}^\top u\right)^2,$$

where $B' = \left\{u : \|u\| = 1,\ u^\top \bar{u} = \left(1 - \frac{1}{2}\varepsilon^2\right)\right\}$. It follows that

$$\min_{u \in B} \left(\bar{u}^\top u\right)^2 = \left(1 - \frac{1}{2}\varepsilon^2\right)^2,$$

whence (39) becomes

$$\bar{u}^\top \Sigma_\infty \bar{u} \leq \beta \left(1 - \left(1 - \frac{1}{2}\varepsilon^2\right)^2\right). \tag{40}$$

Taking $\varepsilon \to 0$ leads to the desired result. □

THEOREM 4. *Suppose that*

$$u_k = \arg \max_{u : \|u\| = 1} \mathbb{K}_\alpha \left(u, \bar{c}_k, \Sigma_k\right),$$

that is, measurements are chosen according to the expected improvement policy. Then, $\mathbb{K}_\alpha(u, \bar{c}_k, \Sigma_k) \to 0$ for all u on the L2 sphere.

Proof. Recall that $\bar{c}_k \to \bar{c}_\infty$ and $\Sigma_k \to \Sigma_\infty$ almost surely. For all u, define $g(u) = \lim_{k \to \infty} \mathbb{K}_\alpha(u, \bar{c}_k, \Sigma_k)$. By Proposition 8, $g(u)$ exists and is finite for all u.

We argue that $\sup_{u : \|u\| = 1} g(u) = 0$. To see this, we write

$$\begin{aligned} \sup_{u : \|u\| = 1} g(u) &= \sup_{u : \|u\| = 1} \liminf_{k \to \infty} \mathbb{K}_\alpha(u, \bar{c}_k, \Sigma_k) \\ &\leq \liminf_{k \to \infty} \sup_{u : \|u\| = 1} \mathbb{K}_\alpha(u, \bar{c}_k, \Sigma_k) \\ &= \liminf_{k \to \infty} \mathbb{K}_\alpha(u_k, \bar{c}_k, \Sigma_k) \\ &\leq \liminf_{k \to \infty} \frac{1}{\sigma_w} \left(\alpha + \frac{2}{\sqrt{2\pi}}\right) \max_{x \in \mathcal{X}} \left|x^\top \Sigma_k u_k\right| \\ &\leq \liminf_{k \to \infty} \frac{1}{\sigma_w} \left(\alpha + \frac{2}{\sqrt{2\pi}}\right) \left(\max_{x \in \mathcal{X}} \sqrt{x^\top \Sigma_0 x}\right) \sqrt{u_k^\top \Sigma_k u_k}. \end{aligned}$$

The third line follows by the definition of u_k. The fourth line follows by Proposition 6. The final line is due to the Cauchy-Schwarz inequality and the monotonicity of the posterior variance. Thus, we have

$$\sup_{u:\|u\|=1} g(u) \leq C \liminf_{k\to\infty} \sqrt{u_k^\top \Sigma_k u_k} \tag{41}$$

for some constant C. We can then take a subsequence $\{u_{k_j}\}$ of $\{u_k\}$ such that $u_{k_j} \to \bar{u}$. By Proposition 10, we know that $\bar{u}^\top \Sigma_\infty \bar{u} = 0$, whence $u_{k_j}^\top \Sigma_{k_j} u_{k_j} \to 0$. Consequently, the right-hand side of (41) is also equal to zero. □

Combining Theorem 4 with Propositions 7 and 8, we find that $u^\top \Sigma_k x \to 0$ for all u on the L2 sphere and all $x \in \mathcal{X}$. It follows that $x^\top \Sigma_k x \to 0$ for all $x \in \mathcal{X}$. That is, asymptotically, we obtain perfect information about the objective value $x^\top c^{\text{true}}$ for any feasible $x \in \mathcal{X}$. This can be viewed as a form of consistency for the policy (as in [52]), in that the policy learns about every decision of interest.

Note that this does not necessarily mean that any u is measured infinitely often (or even once). In fact, it is easy to see from Proposition 10 that $u^\top \Sigma_k u \to 0$ for any u that is in the span of the accumulation points of (u_k), even if u itself is never measured. However, one way or another, we asymptotically obtain perfect information about all relevant x.

9. Numerical tests. Our numerical experiments are implemented in Matlab 7.10. The LPs and SOCPs are solved with the commercial solvers Cplex 12.2.0.2, gurobi or Mosek 7.0.0.64. The SDPs are formulated through cvx [26, 27] in Matlab and then solved with SDPT3 [55] or the commercial solver Mosek 7.

We compare the following algorithms to select measurements when the information state is $(\bar{c}, \Sigma)$:

- EIG: u set to the eigenvector relative to the largest eigenvalue of Σ.
- SDP-1: select u to maximize $\widehat{\mathbb{K}}_\alpha^1(u, \bar{c}, \Sigma)$, using the one-sample approximation scheme of Section 7.1.
- SDP-2: select u to maximize $\widehat{\mathbb{K}}_\alpha^N(u, \bar{c}, \Sigma)$ with $N = 5$, using the general scheme of Section 7.3, with $M = 5$ random positive directions ν_ℓ.
- UNIT: select the unit vector e_i that maximizes $\widehat{\mathbb{K}}_\alpha^N(\cdot, \bar{c}, \Sigma)$ with $N = 21$.
- RAND: select the best random vector u that maximizes $\widehat{\mathbb{K}}_\alpha^N(\cdot, \bar{c}, \Sigma)$ with $N = 21$, among a set of $M = 100$ normalized vectors generated randomly.
- THOMPSON: select the vector $u = \tilde{x}/||\tilde{x}||$, where $\tilde{x}$ is the solution to $\max_{x\in\mathcal{X}} \tilde{c}^\top x - \alpha\sqrt{x^\top \Sigma x}$ with the vector $\tilde{c}$ sampled from $\mathcal{N}(\bar{c}, \Sigma)$.

The algorithms UNIT and RAND are enumeration algorithms that select u out of a finite set of measurements to maximize a good approximation of the expected improvement, $\widehat{\mathbb{K}}_\alpha^N(\cdot, \bar{c}, \Sigma)$ with $N = 21$ in the present case. The policy RAND chooses from a set of $M = 100$ random vectors u generated a priori.

The algorithm THOMPSON adapts the principle of Thompson sampling to our context. Thompson sampling [54] is a randomized algorithm based on the optimization of a problem formed with a single sample from the posterior belief distribution. Thompson sampling has been shown to lead to good empirical and theoretical performance in online learning [10, 1, 47].

In our case, we sample a coefficient $\tilde{c}$ from $\mathcal{N}(\bar{c}_k, \Sigma_k)$, the current belief distribution for c^{true}. We solve a single deterministic SOCP, $\max_{x\in\mathcal{X}} \tilde{c}^\top x - \alpha\sqrt{x^\top \Sigma_k x}$. One issue here is that the solution, say $\tilde{x}$, cannot serve directly as a measurement, since it lives in a different feasibility set. But we can still determine a measurement direction by normalizing $\tilde{x}$, that is, we set $u = x/||\tilde{x}||$. Consequently, this policy can only measure vectors that are scaled versions of feasible implementation decisions.

9.1. Example with robust MDPs. First, we present results obtained on a randomly generated MDP with $|S| = 10$ states and $|A| = 2$ actions. The comparison is done based on the measurement policy induced by the algorithms over a sequence of 10 measurements. We are interested in the true value of the MDP policy that is obtained after k measurements for $k = 1, \ldots, 10$, that is,

$$f(x_k, c^{\text{true}}) = x_k^\top c^{\text{true}} \ , \qquad x_k \in \underset{x:\, Ax=b,\, x\succeq 0}{\arg\max}\ x^\top \bar{c}_k - \alpha\sqrt{x^\top \Sigma_k x} \ ,$$

where $\bar{c}_k$, Σ_k are the end-result of the method that optimizes the measurement vectors $u_1, \ldots, u_k$, and of the random observations $y_1 = u_1^\top c^{\text{true}} + w_1$, …, $y_k = u_k^\top c^{\text{true}} + w_k$.

In the MDP terminology, x_k encodes the stochastic policy π_k that optimally solves the robust MDP (18) posed on the beliefs after k measurements. The vector c^{true} encodes the true reward function $r^{\text{true}}(\cdot,\cdot)$ of the MDP, and $f(x_k, c^{\text{true}})$ is the expected value of π_k on the true MDP, that we can also write as

$$V^{\pi_k} = \mathbb{E}\Big\{ \sum_{t=0}^{\infty} \gamma^t r^{\text{true}}(s_t, \pi_k(s_t)) \,\Big|\, s_0 \sim q_0 \Big\} \ ,$$

for some initial state distribution q_0 determined by the vector b. Because the sequence $w_1, \ldots, w_k$ of observation noise is random, one should actually look at the distribution of $f(x_k, c^{\text{true}}) = V^{\pi_k}$.

Figures 1 to 6 show the results of 100 simulations run on the same fixed MDP. All simulations start from a same belief distribution $(\bar{c}_0, \Sigma_0)$. There are 6 graphs, corresponding to EIG, UNIT, RAND, THOMPSON, SDP-1 and SDP-2. The same 100 samples of a sequence of Gaussian noises $\{w_k : 1 \le k \le 10\}$ for making 10 consecutive measurements are used for comparing the 6 methods. The true maximum is indicated by a horizontal line at 76.59. We have plotted the curve of the estimated mean of V^{π_k} over the 100 samples as a function of the number $k = 0, \ldots, 10$ of past measurements. We have also plotted vertical bars between the 25th and the 75th percentiles of the distribution of V^{π_k}. The support of V^{π_k} cannot cross the horizontal line of the true maximum.

The results show that the two policies proposed in our paper, EIG and SDP-2, generally outperform the other policies. Since our paper considers learning in the context of robust optimization, it is important to look at worst-case performance in addition to the average case. For example, we see that EIG is competitive with SDP-2 on average, but is more irregular in terms of the evolution of the 25th quantile. The policy SDP-1 also performs reasonably well, but exhibits much greater variance compared to SDP-2, illustrating the value added by using $N > 1$ in the SDP framework. Similarly, THOMPSON exhibits steady improvement over time, but likewise exhibits much greater variance compared to SDP-2. It is also noteworthy that the negative tails for SDP-2 are much smaller than the positive tails – the best-case performance is considerably better than the average case, but the worst case is not too much worse.

We believe that the good performance of EIG in this setting can be explained as follows. The terminal value function $v_\alpha(\bar{c}_{10}, \Sigma_{10})$ that we wish to optimize includes a penalty term based on our posterior variance. Because the measurement noise σ_w^2 is known, any measurements will contribute a *deterministic* improvement to the variance term, regardless of the outcome of the observation. Thus, a policy that is designed purely to reduce the uncertainty may actually produce reasonable performance in a risk-averse setting. However, the SDP-2 policy, which considers both the deterministic improvement due to variance reduction and the stochastic improvement coming from the observation, is able to achieve good results faster and reduce the negative tails more consistently.

The computation time to run the experiments over the 100 sample paths in parallel, using 6 processes, are given in Table 1. To get an estimate of the average time to select a single measurement, we divide those numbers by 10, the number of measurements per sample path, and then by (100/6), the number of sample paths divided by the number of processes.

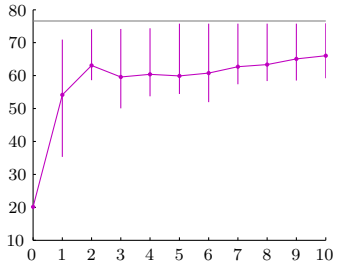

FIGURE 1. Distribution of the true performance with SDP-2, for a growing number of measurements.

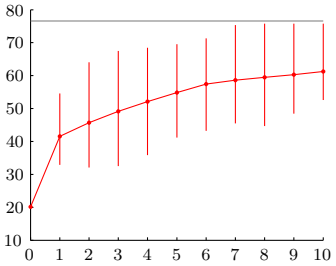

FIGURE 2. Distribution of the true performance with SDP-1, for a growing number of measurements.

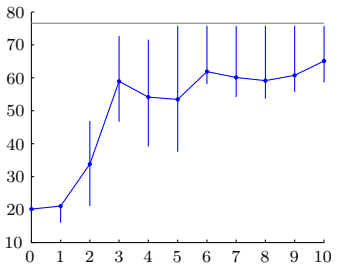

FIGURE 3. Distribution of the true performance with EIG, for a growing number of measurements.

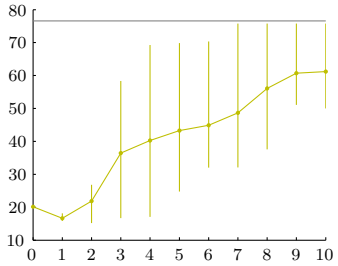

FIGURE 4. Distribution of the true performance with UNIT, for a growing number of measurements.

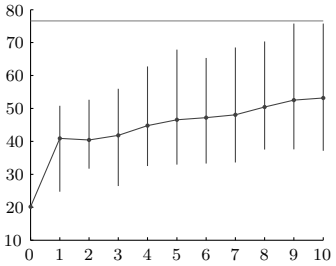

FIGURE 5. Distribution of the true performance with RAND, for a growing number of measurements.

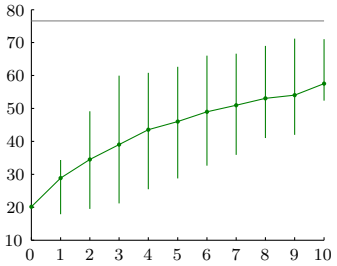

FIGURE 6. Distribution of the true performance with THOMPSON, for a growing number of measurements.

TABLE 1. Computation times (in seconds).

	EIG	UNIT	RAND	THOMPSON	SDP-1	SDP-2
On 100 sample paths:	(1)	(615)	(5849)	(8.5)	(124)	(1239)
Per measurement:	(<0.01)	(3.7)	(35)	(0.05)	(0.74)	(7.4)

9.2. Example with learning in linear regression. Next, we consider a test problem whose structure is motivated by the multi-drug therapy trial problem of [7]. The underlying LP is given by

$$\text{maximize} \quad c^\top x \qquad \text{subject to} \quad a^\top x \leq b \ , \quad 0 \preceq x \preceq x^{\max}.$$

We can view the objective function $c^\top x$ as a prediction of the overall survival of a group of patients, where c is a vector of predicted impacts per unit of each drug, and x is a vector of prescribed dosages. The vector $x^{\max}$ represents the maximum dosage levels allowed for the drugs, while a is a vector of per-unit toxicities, and b is the maximal permissible toxicity in the treatment. Given a belief $c \sim \mathcal{N}(\bar{c}, \Sigma)$, we wish to prescribe dosages that perform well in worst-case scenarios at a certain level of confidence relative to the beliefs. This leads to the robust formulation

$$\text{maximize} \quad c^\top x - \alpha\sqrt{x^\top \Sigma x} \qquad \text{subject to} \quad a^\top x \leq b, \qquad 0 \preceq x \preceq x^{\max} \ .$$

Before deciding on x, we have the ability to collect information about c (e.g. by conducting lab experiments before moving on to trials with human subjects). We model the outcome of such an experiment as $y = c^\top u + w$, where u reflects the weights of the different drugs for the pre-trial study. Essentially the learning process in this problem is an instance of Bayesian linear regression [35]. We assume that $u \succeq 0$ and $\|u\| = 1$. To illustrate the behavior of the model, we consider $n = 40$ drugs with the chosen parameters

$$\bar{c}_i = 3 + \tfrac{i-1}{n-1}, \qquad a_i = 2 + \tfrac{i-1}{n-1}, \qquad x_i^{\max} = 1, \qquad b = 4,$$

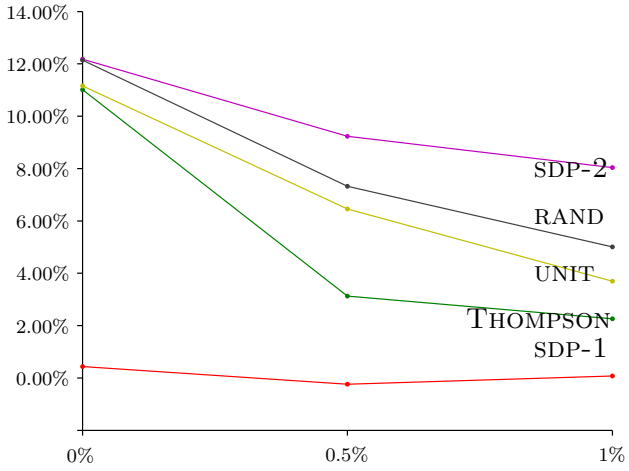

FIGURE 7. Expected improvement, with different risk aversions.

$$\Sigma_{ii} = 0.5 + 1.5\tfrac{i-1}{n-1}, \qquad \Sigma_{ij} = (-1)^{i-j} e^{-2|i-j|} \sqrt{\Sigma_{ii}\Sigma_{jj}} \ .$$

Thus, the drugs go gradually from moderate impact, low variance drugs to higher impact, higher variance drugs. (With the definition of the covariance matrix, the inverse covariance matrix is tridiagonal and nonnegative.) The noise variable w is centered Gaussian with variance 1.

Table 2 describes the distribution of the optimal median objective value, for values of the risk-aversion coefficient $\alpha = 0$ (risk-neutral), $\alpha = 0.5$, and $\alpha = 1.0$. It can be seen that the risk of getting a low true objective value can be decreased, in exchange for a moderate mean reduction, in accordance with the typical behavior of robust optimization models.

Table 3 gives the result of the optimization of the expected improvement in the cases $\alpha = 0, 0.5, 1.0$. As before, the UNIT approach reduces the set of possible studies to those that test one drug at a time (choosing the one with the best expected improvement), while RAND maximizes the expected improvement over a finite set of 1000 randomly generated measurement vectors. The number of sampled vectors was chosen to make RAND run for about as long as the SDP approach. We did not include the EIG policy in this comparison because the dominant eigenvector of Σ may have negative elements, and we require $u \succeq 0$. The THOMPSON policy is randomized, so for it we report the expected improvement averaged over 100 realizations of Thompson sampling.

The maximization of the expected improvement over vectors $u \succeq 0$ with $||u|| = 1$ is performed with the semidefinite programming approach. The constraint $u \succeq 0$ is implemented by adding the constraint $Y_{ij} \geq 0$ for all i, j to the program of Section 7.3, using the property that a nonnegative matrix admits a nonnegative eigenvector. We run the SDP approach with $N = 1$ and $N = 2$ samples only. The computation of the expected improvement for a fixed u (the output of the algorithm) is also done with $N = 21$.

We see in Table 3 that the one-sample approximation (SDP-1) leads to poor results in this setting. In the risk-neutral case ($\alpha = 0$) this is not a complete surprise because the formulation was established under the assumption $\alpha > 0$. On the other hand, the two-sample approximation leads to measurements that dominate those from the other benchmarks, thus illustrating the value added by using multiple samples. Figure 7 demonstrates the dependence of the expected improvement on the risk-aversion parameter α.

Finally, we consider a sequential version of the problem with $K = 10$ measurements and $\alpha = 1$. In the interests of brevity, we only compare SDP-2 and THOMPSON here, as EIG is inapplicable here, and Figure 7 has illustrated that the other policies are much less effective than SDP-2 in optimizing the expected improvement. Figures 8 and 9 report the means and the 25th and 75th

TABLE 2. Distribution of the optimal value of the implementation problem for different risk aversions

	mean	std	percentiles			density
			25th	10th	5th	
$\alpha = 0$	5.99	0.27	5.70	5.63	5.53	
$\alpha = 0.5$	5.96	0.15	5.80	5.77	5.71	
$\alpha = 1$	5.94	0.12	5.81	5.78	5.74	

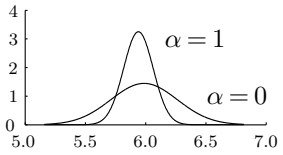

TABLE 3. Value of the expected improvement for different risk aversions and optimization approaches

	UNIT			RAND			THOMPSON			SDP-2		
	$\mathbb{K}_\alpha$		time	$\mathbb{K}_\alpha$		time	$\mathbb{K}_\alpha$		time	$\mathbb{K}_\alpha$		time
$\alpha = 0$	0.67	11.2%	(33)	0.73	12.2%	(644)	0.66	11.0%	(20)	0.73	12.19%	(627)
$\alpha = 0.5$	0.38	6.43%	(49)	0.43	7.28%	(991)	0.18	3.12%	(19)	0.55	9.31%	(556)
$\alpha = 1$	0.21	3.61%	(49)	0.29	4.98%	(990)	0.13	2.26%	(20)	0.47	8.78%	(511)

The expected improvement is given in absolute units, and in percentage of the optimal value of the program. Computation times in seconds, using 6 parallel processes. SDP-1 gives in all cases an expected improvement close to 0 and is omitted in this table.

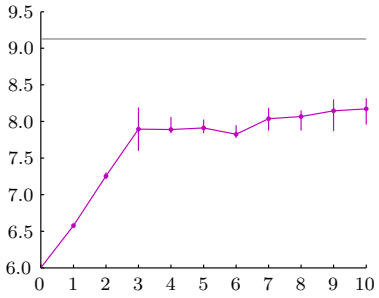

FIGURE 8. Distribution of the true performance with SDP-2, for a growing number of measurements.

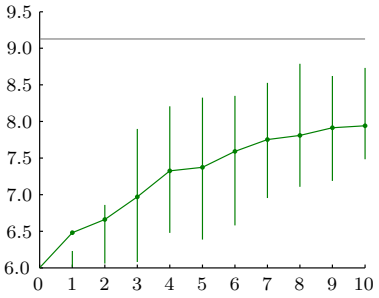

FIGURE 9. Distribution of the true performance with THOMPSON, for a growing number of measurements.

percentiles of the distributions of the implementation decisions for the two policies. We see that SDP-2 outperforms THOMPSON on average, but much more importantly, the performance of SDP-2 exhibits much smaller variation. This shows the efficacy of the method in a risk-averse setting.

Note that when there is no measurement (origin of the horizontal axis), the implementation decision x_0 is in all cases selected as the maximizer of $\bar{c}^\top x - \alpha\sqrt{x^\top \Sigma x}$. This leads to a single value for the performance with no measurements, namely $c^{\text{true}\top} x_0$. This value can be viewed as being sampled from the distribution reported in the row relative to $\alpha = 1$ in Table 3.

10. Conclusion. We have posed an optimal learning problem in which a decision-maker improves a robust solution to a stochastic linear program by sequentially collecting information about the unknown objective coefficients. A single piece of information takes the form of a linear combination (a "blend") of the true underlying objective vector, subject to Gaussian noise. Bayesian updating is then used to combine this new information with a multivariate normal prior distribution on the unknown parameters. Previous work has considered weighted sums of unknown parameters where the weights were pre-specified by a linear regression model. To our knowledge,

the present paper is the first to pose the continuous optimization problem of choosing the optimal weight vector. Our formulation of this problem allows for both risk-neutral and risk-averse decision-makers.

Within this setting, we have proposed two policies for choosing information blends. The first was shown to optimize uncertainty reduction (analogous to active learning methods in statistics) by selecting the largest eigenvector of the posterior covariance matrix. The second approximates the optimal solution to an expected improvement criterion (a nonconvex optimization problem) via an SDP reformulation technique. The approach is applicable to robust LP formulations of Markov decision process problems, where risk-averse decision-making policies are desired. We show that our approach generalizes a previous heuristic for such problems. In numerical examples, the SDP approximation consistently outperforms a number of benchmarks. We believe that the present paper contributes to the interface of robust optimization and optimal learning, and that the idea of information blending offers a new way to think about sequential information collection.

Appendix. Proof of Proposition 1.

Assuming $\alpha > 0$, we have, from (19),

$$\arg\max_{u\in\mathbb{B}} \widetilde{\mathbb{K}}_\alpha(u,\bar{c},\Sigma) = \arg\max_{u\in\mathbb{B}} \sqrt{\bar{x}^\top\Sigma\bar{x}} - \sqrt{\bar{x}^\top\left(\Sigma - \frac{\Sigma uu^\top\Sigma}{u^\top\Sigma u+\sigma_w^2}\right)\bar{x}} \ .$$

If $\Sigma\bar{x}=0$, then any $u\in\mathbb{B}$ is optimal. Otherwise, $\Sigma\bar{x}\neq 0$, and we can justify that any optimal u will satisfy $u^\top u=1$ by the proof technique used in Theorem 1. Then we have

$$\begin{aligned}
\arg\max_{u\in\mathbb{B}} \widetilde{\mathbb{K}}_\alpha(u,\bar{c},\Sigma) &= \arg\max_{u:\,||u||=1} \sqrt{\bar{x}^\top\Sigma\bar{x}} - \sqrt{\bar{x}^\top\left(\Sigma - \frac{\Sigma uu^\top\Sigma}{u^\top\Sigma u+\sigma_w^2}\right)\bar{x}} \\
&= \arg\min_{u:\,||u||=1} \sqrt{\bar{x}^\top\left(\Sigma - \frac{\Sigma uu^\top\Sigma}{u^\top\Sigma u+\sigma_w^2}\right)\bar{x}} \\
&= \arg\min_{u:\,||u||=1} \bar{x}^\top\left(\Sigma - \frac{\Sigma uu^\top\Sigma}{u^\top\Sigma u+\sigma_w^2}\right)\bar{x} \\
&= \arg\max_{u:\,||u||=1} \frac{\bar{x}^\top\Sigma uu^\top\Sigma\bar{x}}{u^\top\Sigma u+\sigma_w^2} \\
&= \arg\max_{u:\,||u||=1} \frac{u^\top\Sigma\bar{x}\bar{x}^\top\Sigma u}{u^\top(\Sigma+\sigma_w^2\mathrm{I}_n)u} \ .
\end{aligned}$$

We can then proceed as in the proof of Theorem 2, or observe that an optimal solution $\bar{u}$ can be obtained by considering the generalized eigenvalue problem $(\Sigma\bar{x}\bar{x}^\top\Sigma)u=\lambda(\Sigma+\sigma_w^2\mathrm{I}_n)u$ and taking for $\bar{u}$ a normalized generalized vector associated to the largest generalized eigenvalue λ. Since $(\Sigma+\sigma_w^2\mathrm{I}_n)$ is nonsingular, the generalized eigenvalue problem is equivalent to the standard eigenvalue problem $(\Sigma+\mathrm{I}_n\sigma_w^2)^{-1}(\Sigma\bar{x}\bar{x}^\top\Sigma)u=\lambda u$, which is of the form

$$fg^\top u=\lambda u \quad \text{with } f=(\Sigma+\mathrm{I}_n\sigma_w^2)^{-1}\Sigma\bar{x} \ , \quad g=\Sigma\bar{x}.$$

Therefore, the rank-one matrix $fg^\top$ has a single positive eigenvalue $g^\top f/||f||$ with a normalized eigenvector $f/||f||$ or $-f/||f||$, and $\bar{u}=\pm(\Sigma+\mathrm{I}_n\sigma_w^2)^{-1}\Sigma\bar{x}/||(\Sigma+\mathrm{I}_n\sigma_w^2)^{-1}\Sigma\bar{x}||$.

References

[1] Agrawal, S., N. Goyal. 2013. Thompson sampling for contextual bandits with linear payoffs. *Proc. 30th Internat. Conf. Machine Learning (ICML-2013)*. 337–344.

[2] Alizadeh, F., D. Goldfarb. 2003. Second-order cone programming. *Math. Programming* **95** 3–51.

[3] Auer, P., N. Cesa-Bianchi, P. Fischer. 2002. Finite-time analysis of the multiarmed bandit problem. *Machine Learning* **47** 235–256.

[4] Barvinok, A. 2001. A remark on the rank of positive semidefinite matrices subject to affine constraints. *Discrete and Computational Geometry* **25** 23–31.

[5] Ben-Tal, A., L. El Ghaoui, A. Nemirovski. 2009. *Robust Optimization*. Princeton University Press.

[6] Ben-Tal, A., A. Goryashko, E. Guslitzer, A. Nemirovski. 2004. Adjustable robust solutions of uncertain linear programs. *Math. Programming* **99**(2) 351–376.

[7] Bertsimas, D., A. O'Hair, S. Relyea, J. Silberholz. 2013. An analytics approach to designing clinical trials for cancer. *Working paper, MIT* URL http://josilber.scripts.mit.edu/CancerPaper_Revision1_names.pdf.

[8] Bertsimas, D., M. Sim. 2004. The price of robustness. *Oper. Res.* **51**(1) 35–53.

[9] Bubeck, S., N. Cesa-Bianchi, S. M. Kakade. 2012. Towards minimax policies for online linear optimization with bandit feedback. *Proc. 25th Conf. Learning Theory (COLT-2012)*.

[10] Chapelle, O., L. Li. 2011. An empirical evaluation of Thompson sampling. *Adv. in Neural Inform. Processing Systems 24 (NIPS-2011)*. 2249–2257.

[11] Chick, S. E. 2006. Subjective probability and Bayesian methodology. S. G. Henderson, B. L. Nelson, eds., *Handbooks of Operations Research and Management Science, vol. 13: Simulation*. North-Holland Publishing, Amsterdam, 225–258.

[12] Chick, S. E., N. Gans. 2009. Economic analysis of simulation selection problems. *Management Sci.* **55**(3) 421–437.

[13] Cohn, D.A., Z. Ghahramani, M.I. Jordan. 1996. Active learning with statistical models. *J. Artificial Intelligence Res.* **4** 129–145.

[14] Dani, V., T. P. Hayes, S. M. Kakade. 2008. Stochastic linear optimization under bandit feedback. *Proc. 21st Conf. Learning Theory (COLT-2008)*. 355–366.

[15] Delage, E., S. Mannor. 2007. Percentile optimization in uncertain Markov decision processes with application to efficient exploration. *Proc. 24th Internat. Conf. Machine Learning (ICML-2007)*. ACM, 225–232.

[16] Delage, E., S. Mannor. 2010. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Oper. Res.* **58**(1) 203–213.

[17] Dellino, G., J.P.C. Kleijnen, C. Meloni. 2012. Robust optimization in simulation: Taguchi and Krige combined. *Informs J. Comp.* **24**(3) 471–484.

[18] D'Epenoux, F. 1960. Sur un problème de production et de stockage dans l'aléatoire. *Revue Française de Recherche Opérationnelle* **14** 3–16.

[19] Dontchev, A.L., R.T. Rockafellar. 2009. *Implicit Functions and Solution Mappings*. Springer.

[20] Doob, J.L. 1949. Application of the theory of martingales. *Le calcul des probabilités et ses applications*. Colloques Internationaux du Centre National de la Recherche Scientifique, no. 13, CNRS, Paris, 23–27.

[21] Fazel, M., H. Hindi, S. Boyd. 2001. A rank minimization heuristic with application to minimum order system approximation. *Proc. 2001 American Control Conference*. Arlington, VA, 4734 – 4739.

[22] Frazier, P.I., W.B. Powell, S. Dayanik. 2008. A knowledge gradient policy for sequential information collection. *SIAM J. Control Optim.* **47**(5) 2410–2439.

[23] Ghaffari-Hadigheh, A., T. Terlaky. 2006. Sensitivity analysis in linear optimization: Invariant support set intervals. *Eur. J. Oper. Res.* **169** 1158–1175.

[24] Gittins, J. C., K. D. Glazebrook, R. Weber. 2011. *Multi-armed bandit allocation indices*. 2nd ed. John Wiley and Sons, Chichester, UK.

[25] Graf, S., H. Luschgy. 2000. *Foundations of Quantization for Probability Distributions*. Springer-Verlag, Berlin, Germany.

[26] Grant, M., S. Boyd. 2008. Graph implementations for nonsmooth convex programs. V. Blondel, S. Boyd, H. Kimura, eds., *Recent Advances in Learning and Control – A tribute to M. Vidyasagar*. LNCIS, Springer, New York, NY, 95–110.

[27] Grant, M., S. Boyd. 2011. CVX: Matlab software for disciplined convex programming, version 1.21. http://cvxr.com/cvx.

[28] Gupta, S.S., K.J. Miescke. 1996. Bayesian look ahead one-stage sampling allocations for selection of the best population. *J. Statist. Planning and Inference* **54**(2) 229–244.

[29] Han, B., I. O. Ryzhov, B. Defourny. 2013. Efficient learning of donor retention strategies for the American Red Cross. R. Pasupathy, S.-H. Kim, A. Tolk, R. Hill, M. E. Kuhl, eds., *Proc. 2013 Winter Simulation Conf.*.

[30] Iyengar, G. N. 2005. Robust dynamic programming. *Math. Oper. Res.* **30**(2) 257–280.

[31] Jones, D.R., M. Schonlau, W.J. Welch. 1998. Efficient global optimization of expensive black-box functions. *J. Global Optim.* **13**(4) 455–492.

[32] Kim, S.-H., B. L. Nelson. 2006. Selecting the best system. S. G. Henderson, B. L. Nelson, eds., *Handbooks of Operations Research and Management Science, vol. 13: Simulation*. North-Holland Publishing, Amsterdam, 501–534.

[33] Mangasarian, O.L., T.H. Shiau. 1986. A variable-complexity norm maximization problem. *SIAM J. Algebraic Discrete Methods* **7**(3) 455–461.

[34] McMahan, H. B., G. J. Gordon, A. Blum. 2003. Planning in the presence of cost functions controlled by an adversary. *Proc. 20th Internat. Conf. Machine Learning (ICML-2003)*. 536–543.

[35] Minka, T. P. 2000. Bayesian linear regression. Tech. rep., Microsoft Research.

[36] Negoescu, D. M., P. I. Frazier, W. B. Powell. 2010. The knowledge-gradient algorithm for sequencing experiments in drug discovery. *Informs J. Comp.* **23**(3) 346–363.

[37] Nesterov, Y., H. Wolkowicz, Y. Ye. 2000. Semidefinite programming relaxations of nonconvex quadratic optimization. H. Wolkowicz, R. Saigal, L. Vandenberghe, eds., *Handbook of Semidefinite Programming*. Springer.

[38] Nilim, A., L. El Ghaoui. 2005. Robust control of Markov decision processes with uncertain transition matrices. *Oper. Res.* **53**(5) 780–798.

[39] Pages, G., J. Printems. 2003. Optimal quadratic quantization for numerics: the Gaussian case. *Monte Carlo Methods and Applications* **9**(2) 135–166.

[40] Pataki, G. 1998. On the rank of extreme matrices in semidefinite programs and the multiplicity of optimal eigenvalues. *Math. Oper. Res.* **23**(2) 339–358.

[41] Powell, W. B. 2011. *Approximate dynamic programming: solving the curses of dimensionality (2nd ed.)*. John Wiley and Sons.

[42] Powell, W. B., I. O. Ryzhov. 2012. *Optimal Learning*. Wiley, Hoboken, NJ.

[43] Puterman, M.L. 1994. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley, New York, NY.

[44] Regan, K., C. Boutilier. 2009. Regret-based reward elicitation for Markov decision processes. *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*. 444–451.

[45] Regan, K., C. Boutilier. 2010. Robust policy computation in reward-uncertain MDPs using nondominated policies. *Proc. 24th AAAI Conf. Artificial Intelligence (AAAI-10)*. The AAAI Press, Menlo Park, CA, 1127–1133.

[46] Rockafellar, R.T. 1970. *Convex Analysis*. Princeton University Press, Princeton, NJ.

[47] Russo, D., B. Van Roy. 2013. Learning to optimize via posterior sampling. *arXiv preprint arXiv:1301.2609* .

[48] Ruszczyński, A. 2010. Risk-averse dynamic programming for Markov decision processes. *Math. Programming* **125**(2) 235–261.

[49] Ryzhov, I. O., B. Defourny, W. B. Powell. 2012. Ranking and selection meets robust optimization. *Proc. 2012 Winter Simulation Conference*.

[50] Ryzhov, I.O., W.B. Powell. 2011. Information collection on a graph. *Oper. Res.* **59**(1) 188–201.

[51] Ryzhov, I.O., W.B. Powell. 2012. Information collection for linear programs with uncertain objective coefficients. *SIAM J. Optim.* **22**(4) 1344–1368.

[52] Scott, W. R., P. I. Frazier, W. B. Powell. 2011. The correlated knowledge gradient for simulation optimization of continuous parameters using Gaussian process regression. *SIAM J. Optim.* **21**(3) 996–1026.

[53] Shor, N.Z. 1987. Quadratic optimization problems. *Soviet Journal of Circuits and Systems Sciences* **25** 1–11.

[54] Thompson, W.R. 1933. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* **25**(3–4) 285–294.

[55] Tütüncü, R.H., K.C. Toh, M.J. Todd. 2003. Solving semidefinite-quadratic-linear programs using SDPT3. *Math. Programming* **95** 189–217.

[56] Waeber, R., P. I. Frazier, S. G. Henderson. 2010. Performance measures for ranking and selection procedures. B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan, E. Yücesan, eds., *Proc. 2010 Winter Simulation Conference*. 1235–1245.

[57] Xu, H., S. Mannor. 2009. Parametric regret in uncertain Markov decision processes. *Proc. 48th IEEE Conf. Decision and Control*. 3606–3613.

[58] Zheng, X.J., X.L. Sun, D. Li. 2011. Convex relaxations for nonconvex quadratically constrained quadratic programming: matrix cone decomposition and polyhedral approximation. *Math. Programming, Ser. B* **129**(2) 301–329.